McGRAW-HILL
PUBLICATIONS
IN PSYCHOLOGY

McGRAW-HILL PUBLICATIONS IN PSYCHOLOGY
CLIFFORD T. MORGAN, Consulting Editor

FUNDAMENTAL STATISTICS IN PSYCHOLOGY AND EDUCATION

McGraw-Hill Publications in Psychology

CLIFFORD T. MORGAN

CONSULTING EDITOR

Barker, Kounin, and Wright—CHILD BEHAVIOR AND DEVELOPMENT
Brown—PSYCHOLOGY AND THE SOCIAL ORDER
Brown—THE PSYCHODYNAMICS OF ABNORMAL BEHAVIOR
Cattell—PERSONALITY
Cole—GENERAL PSYCHOLOGY
Crafts, Schneirla, Robinson, and Gilbert—RECENT EXPERIMENTS IN PSYCHOLOGY
Davis—PSYCHOLOGY OF LEARNING
Dorcus and Jones—HANDBOOK OF EMPLOYEE SELECTION
Dunlap—RELIGION: ITS FUNCTIONS IN HUMAN LIFE
Ghiselli and Brown—PERSONNEL AND INDUSTRIAL PSYCHOLOGY
Gray—PSYCHOLOGY IN HUMAN AFFAIRS
Guilford—FUNDAMENTAL STATISTICS IN PSYCHOLOGY AND EDUCATION
Guilford—PSYCHOMETRIC METHODS
Hurlock—ADOLESCENT DEVELOPMENT
Hurlock—CHILD DEVELOPMENT
Johnson—ESSENTIALS OF PSYCHOLOGY
Krech and Crutchfield—THEORY AND PROBLEMS OF SOCIAL PSYCHOLOGY
Lewin—A DYNAMIC THEORY OF PERSONALITY
Lewin—PRINCIPLES OF TOPOLOGICAL PSYCHOLOGY
Maier—FRUSTRATION
Maier and Schneirla—PRINCIPLES OF ANIMAL PSYCHOLOGY
Miller—EXPERIMENTS IN SOCIAL PROCESS
Moore—PSYCHOLOGY FOR BUSINESS AND INDUSTRY
Morgan and Stellar—PHYSIOLOGICAL PSYCHOLOGY
Page—ABNORMAL PSYCHOLOGY
Pillsbury—AN ELEMENTARY PSYCHOLOGY OF THE ABNORMAL
Reymert—FEELINGS AND EMOTIONS
Richards—MODERN CLINICAL PSYCHOLOGY
Seashore—PSYCHOLOGY OF MUSIC
Seward—SEX AND THE SOCIAL ORDER
Stagner—PSYCHOLOGY OF PERSONALITY
Wallin—PERSONALITY MALADJUSTMENTS AND MENTAL HYGIENE

John F. Dashiell was Consulting Editor of this series from its inception in 1931 until January 1, 1950.

Fundamental Statistics in Psychology and Education

BY

J. P. GUILFORD

Professor of Psychology, University of Southern California

SECOND EDITION
SECOND IMPRESSION

McGRAW-HILL BOOK COMPANY, INC.

NEW YORK TORONTO LONDON

1950

FUNDAMENTAL STATISTICS IN PSYCHOLOGY AND EDUCATION

THE MAPLE PRESS COMPANY, YORK, PA.

Keep close to experience; add as little of your own as possible; if you have to add something, be mindful to give an account of every step you take.—F. M. URBAN.

PREFACE TO THE SECOND EDITION

Seven eventful years have elapsed since the appearance of the first edition of this volume. So extensive have been the changes in research and instruction and so rapid have been the developments in statistical theory and method that the volume has been virtually rewritten. With greater recognition of the fact that a textbook in this field is inevitably forced into use as a handbook, many additional topics and features have been added. Many users of the first edition in classroom instruction have kindly made helpful suggestions. No topic was recommended for omission, but the author was urged to include many new ones. The consequence is a much enlarged volume. The emphasis remains upon applications rather than upon mathematical statistics.

Among the topics receiving greater attention are those concerned with categorical data, sampling statistics, analysis of variance, prediction, reliability and validity of tests, and scaling procedures. Methods for tabular and graphic treatment of categorical data are introduced. Standard errors of statistics obtained under varieties of sampling conditions are provided. More attention is given to the logic of statistical inferences and to small-sample statistics. Distributions of t, chi-square, and of the F ratio are illustrated. A two-way classification problem in analysis of variance is described and illustrated. New methods of prediction of attributes are presented and also procedures for evaluating predictions, with tests of significance. Principles of multiple prediction are more clearly brought out and alternative procedures presented. Attention is given to the variances and correlations of weighted and unweighted composites.

Recent developments in the rationale underlying test procedures are reflected in a systematic manner. The different meanings and types of reliability and validity are delineated. Factor theory is now regarded as essential for the understanding of such phenomena as intercorrelation, multiple prediction, and correction for attenuation. A brief introduction is therefore given to factor theory. A description of factor-analysis procedures, however, is much too demanding of space for inclusion in a volume of this character. Scaling procedures have been moved to a final chapter, and the method of paired comparisons has been included.

Throughout, much greater care has been taken to attempt to give the

mathematically unsophisticated student an appreciation of the underlying logic of statistics. Assumptions underlying methods are rather fully mentioned. An appendix has been added in which many simple, yet uncommon, proofs and derivations are presented. The limitations and peculiarities of each statistic are emphasized. It is hoped that all these measures will help to guide the reader in the proper use of statistics and in the avoidance of blunders. Many interrelations among different statistics, heretofore usually unrealized or ignored, have been pointed out. This should bring out for the student a much more coherent picture than before. Many points are liberally repeated in varied form, for emphasis and for evaluative reasons.

A slight rearrangement of the order of chapters now places all the material that is likely to be needed in a first-semester course in statistics first in the volume. This material includes Chaps. 1 through 8 and parts of 9. The remainder includes more than enough material for a second semester's work.

The author is much indebted to the many users of the first edition who responded to requests for criticisms. Among these must be mentioned Dr. William B. Michael and Dr. Harrison F. Heath, who made numerous useful suggestions. Dr. Michael has read a number of the chapters in manuscript form. To the Army Air Forces, from which source many a useful illustration has been drawn, the author owes very much. It has been impossible to mention by name all the individuals who originated ideas (which have been borrowed to great advantage) as a part of their more or less anonymous contribution to the AAF Aviation Psychology Research Program. The author feels sure that none of the information revealed from that source will reflect anything but credit upon the AAF's enlightened personnel-research program during the Second World War. Acknowledgments for the use of other specific materials are given at appropriate places. The author is very much indebted to his wife, Ruth B. Guilford, and to Mrs. William W. Burke for considerable editorial assistance in the preparation of the manuscript.

J. P. Guilford

Beverly Hills, Calif.
December, 1949

PREFACE TO THE FIRST EDITION

Since the publication of *Psychometric Methods* six years ago, the author has sensed a growing need for a supplementation in the form of a textbook on statistical method. This is so because the earlier volume emphasized methods of measurement rather than statistics and because even in the short span of years since its publication, some marked changes of emphasis and some innovations have occurred within psychological and educational statistics. The present volume therefore attempts to conserve what is most useful among the old and to introduce as much of the new as seems to have contemporary and future value.

The treatment presupposes no previous study of statistics and so strives to provide for the student a simple introduction. The more fundamental and useful procedures are outlined step by step and are fully illustrated. The selection of what seems most useful to the present student and investigator has been aided by the author's experience in guiding students in research and in serving as director of the Bureau of Instructional Research at the University of Nebraska. His experiences with teaching the subject over a period of years have dictated the mode of presentation of statistical ideas and methods, but he realizes that sometimes what seem the best modes of presentation are none too good and that much teaching remains to be done. As a textbook, the volume will serve either for a one-semester introductory course or for a full-year course.

Among the innovations in this type of text will be found several features. Some of the graphic devices for representing data are new to textbooks. The treatment of centile norms and of profiles based upon centiles is, the author believes, an improvement. A *C*-scale procedure for normalizing and standardizing scores is proposed and described, along with the traditional *T*-scaling method. The growing emphasis upon sampling and drawing inferences about populations from sample statistics is reflected. Small-sample statistics, including the *t* ratio and Student's distribution are given extensive application. An introduction to analysis of variance is provided, and a novel, pedagogically simple derivation of the analysis-of-variance principle is presented. In a chapter quite new to this type of text, entitled Testing Hypotheses, much of Fisher's work is reflected, and chi square is given prominence. In another new chapter, entitled Predictions and Errors of Prediction, some new devices of practical importance

are introduced. In this chapter and in others, much attention is given to enumeration data and the statistics of attributes, a field that is growing in importance in the social sciences in general. A treatment of factor analysis has been omitted for the reasons that this subject cannot any longer be adequately presented in the space that a text of these proportions would permit, and its study and mastery extend well beyond the student's first year of statistics. Even the final chapter on Mental Tests had to be treated rather sketchily in order to remain within reasonable bounds of space allotted to the volume. In general, there was recognition of the limitations, self-imposed, where references to more advanced treatments were regretfully made in order to stay within the bounds of a fundamental statistics.

The author gladly expresses acknowledgments and thanks to Prof. Harry Helson, who read and criticized three of the chapters. To H. M. Cox, with whom the author was associated in the Bureau of Instructional Research, he owes much for certain ideas regarding ways of presentation of data and concerning the selection of useful methods. To his wife, Ruth B. Guilford, the author is, as always, most indebted for constant help in the preparation of the manuscript. To publishers and authors who have generously permitted the reproduction or use of material he is grateful. These and other contributions are acknowledged specifically at various places in the volume. To Prof. R. A. Fisher and to Messrs. Oliver & Boyd of Edinburgh the author is indebted for permission to reprint Table E from their book *Statistical Methods for Research Workers*, 8th ed., 1942.

J. P. GUILFORD

SANTA ANA, CALIF.
September, 1942

CONTENTS

CHAPTER 1

INTRODUCTION FOR STUDENTS

This book was written for students. It was written by one who has known many students, particularly in psychology and education. He has seen many generations of students prepare for their professions and later at work in their professions. He has seen them take early steps in mastering the methods of their professions. He has seen them at work in the laboratory, in the clinic, and later in the industrial and military laboratory of research and the personnel office. There has been ample opportunity in these experiences to see what the student needs in the way of statistics, and why.

Why the Student Needs Statistics.—Most seasoned workers in psychology or in education usually take the statistical methods for granted as an essential part of their routine; some more so, and some less. The initiate may at first react to statistics as a frightful bogie whose mysteries loom forbidding before him, and he is likely to ask, "What is the good of them, anyway?" This is particularly true of one who feels he has always had trouble with numbers. Students who enter a first course in statistical method in psychology or education, and probably in all related social sciences, range all the way from those who find mathematics in general easy and to their liking, to those at the other extreme who say they have difficulty in adding two and two. Somehow, all of these must acquire what they can of a subject for which they are so unequally prepared.

Probably no other subject demonstrates so clearly that there are several kinds of intelligence. No less a person intellectually than Charles Darwin had trouble with statistics, as he is said to have frankly admitted. His almost equally illustrious cousin, Sir Francis Galton, who is believed to have had an *IQ* of about 200, and who had so much to do with introducing statistics into psychology, had to turn some of his mathematical problems over to others for aid.

There are different ways of understanding the same things. One student will grasp the new ideas offered by statistics in the way that a mathematician would understand them; another will appreciate the logical rules of thinking and the concepts provided as aids in thinking; still others will master rule-of-thumb operations and be able to carry through

computations with a minimum grasp of what they are all about. Learning without achieving insights and appreciations of the inner nature of things is learning without full motivation and enthusiasm, and is not very satisfying. The average student will necessarily have to be content with levels of insight that fall short of those of the mathematician, remembering that even mathematicians have not by any means exhausted the meanings and ramifications of statistical ideas. On the other hand, each student should strive to inject as much meaning and significance, in his own way, as he can. The proper use and optimal use of statistical methods and statistical thinking require certain minimal achievement of understanding. Clerks can be taught to carry out many of the computational procedures; it is not the primary purpose of this book or of those who teach with it to develop computational clerks. The purpose is to develop those who could be supervisors of clerks.

To be more specific, there are four simple, undeniable reasons why the student who takes a required course in statistics must develop some mastery of that subject.

1. *He must be able to read professional literature.*—There is no questioning the fact that learning in any field comes largely through reading. The student never finishes the extension of his skill in the art of reading, if he is a successful student. In any specialized field, reading is largely a matter of enlarging vocabulary. One cannot read much of the literature in any specialized field in the social sciences, particularly psychology and education, without encountering statistical symbols, concepts, and ideas on every hand. One could do as the young child does when he tackles reading matter that is somewhat beyond him, "skip over the hard places." But this is hardly excusable in the adult who is reading material that should not be beyond him and in which the "hard places" may, in fact, contain the crucial parts of the content. One who dodges such parts is likely to be dependent upon the conclusions of others for his own conclusions and opinions. This is hardly independent judgment or a symptom of mature scholarship. It is not necessary for every psychologist to be able to sail through the "heavier" mathematical contributions of the specialist in statistics. It is severely limiting, however, for a person not to be able to read intelligently the average research paper in his field with some appreciation as to whether sound conclusions have been reached. The chances are that this appreciation will require familiarity with the basic statistical ideas.

2. *He must master techniques needed in advanced courses.*—Whether the advanced course is a laboratory course or a practicum, there are usually certain incidental techniques that are commonly used in the operations

involved. In the laboratory course, results cannot be treated or reports written without at least minimal statistical operations. A field survey or the checking of a report also involves inevitable statistical steps.

3. *Statistics is an essential part of professional training.*—The trained psychologist or educator likes to think of himself as a professional person. To some extent, statistical logic, statistical thinking, and statistical operations are a necessary part of either profession. To the extent that he uses in his practice the common technical instruments, such as tests, the psychologist or educator will depend upon statistical background in their administration and in the interpretation of the results. Using tests without knowledge of the statistical reasoning upon which they depend is like the medical diagnostician's using clinical tests without a knowledge of physiology and pathology.

4. *Statistics are everywhere basic to research activities.*—To the extent that either psychologist or educator intends to keep alive his research interests and research activities, he will necessarily lean upon his knowledge and skills in statistical methods. The relation of statistics to research will be elaborated upon in the next paragraphs. Here it is merely urged that in any professional fields where there are still so many unknowns as in psychology and education, the advancement of those professions and of the competence of their members depends to a high degree upon the continued research attitude and research efforts of those members.

Why Statistics Are Important in Research.—Briefly, the advantages of statistical thinking and operations in research are as follows:

1. *They permit the most exact kind of description.*—When all is said and done, the goal of science is description of phenomena, description so complete and so accurate that it is useful to anyone who can understand it when he reads the symbols in terms of which those phenomena are described. Mathematics and statistics are a part of our descriptive language, an outgrowth of our verbal symbols, peculiarly adapted to the efficient kind of description that the scientist demands.

2. *They force us to be definite and exact in our procedures and in our thinking.*—The writer once heard a prominent psychologist defend his rather vague conclusions by saying that he would rather be vague and right than to be definite and wrong. But the alternatives are not to be either "vague and right" or "definite and wrong." One can also be definite and right, and it is the writer's contention that the odds for being right are overwhelmingly on the "definite" side of the matter.

3. *Statistics enable us to summarize our results in meaningful and convenient form.*—Masses of observations taken by themselves are bewilder-

ing and almost meaningless. Before we can see the forest as well as the trees, order must be given to the data. Statistics provide an unrivaled device for bringing order out of chaos; of seeing the general picture in one's results.

4. *They enable us to draw general conclusions*, and the process of extracting conclusions is carried out according to accepted rules. Furthermore, by means of statistical steps, we can say about how much faith should be placed in any conclusion and about how far we may extend our generalization.

5. *They enable us to make predictions* of "how much" of a thing will happen under conditions we know and have measured. For example, we can predict the probable mark a freshman will earn in college algebra if we know his score in a general scholastic-ability test, his score in a special algebra-aptitude test, his average mark in high-school mathematics, and perhaps the number of hours per week that he devotes to studying algebra. Our prediction may be somewhat in error because of other factors that we have not accounted for, but our statistical methods will also tell us about how much margin of error to allow in our predictions. Thus not only can we make predictions but we know how much faith to place in them.

6. *They enable us to analyze some of the causal factors out of complex and otherwise bewildering events.*—It is generally true in the social sciences, and in psychology and education in common with them, that any event or outcome is a resultant of numerous causal factors. The reasons why a man fails in his business or in his profession, for example, are varied and many. Causal factors are usually best uncovered and proved by means of experimental method. If it could be shown that, all other factors being held constant, certain business men fail to the extent that they possess some defect of personality "X," then it is probable that X is a cause of failure in this type of business. Unfortunately for the social scientist, he cannot manage men and their affairs sufficiently to set up a good experiment of this type. The next best thing is to make a statistical study, taking business men as we find them, working under conditions as they normally do. The life-insurance expert does the same kind of thing when he follows the trail of all possible factors that influence the length of life and determines how important they are. On the basis of these statistical findings, he can predict about how long an individual of a certain type will probably live, and his insurance company can plan an insurance policy accordingly. Statistical methods are therefore often a necessary substitute for experiments. Even where experiments are possible, the experimental data must ordinarily receive appropriate statistical

treatment. Statistical methods are hence the constant companions of experiments.

What This Volume's Treatment of Statistics Will Include.—For the next few paragraphs we will take a hasty overview of the things to come. The second chapter will give many more details of a general and preparatory nature. Here we will try to look at the whole forest before we enter it.

Descriptive and Sampling Statistics.—With some writers it is common to make a broad distinction between descriptive and sampling statistics. The distinction is real and should be realized, although we need not be perpetually aware of it. It refers to two important uses of statistics. In the first place, statistics are used to describe situations. Averages tell us "how much" of certain quantities we have in a group of individuals or in a group of observations. An average (*e.g.*, *arithmetic mean*, *median*, or *mode*) is a general-level concept. A single number tells how high one group, or sample, stands on a certain scale as compared with another.

Other statistics tell us how much variability or scatter the individuals of a group show. A statistic known as the *standard deviation* has been the almost universal indicator of the amount of variability in a set of individuals or observations, though there are others.

A *coefficient of correlation* describes the closeness of relationship between two sets of measures of the same group of individuals or observations. Most of science is concerned in finding out what things go with what, and what things are independent of what. Correlation methods, in the social sciences at least, are the most useful devices to answer these questions of interrelationships. Averages, indices of dispersion and of correlation, are the basic and chief descriptive statistics.

Sampling statistics have become increasingly important in recent years. Their use is to tell us how well the statistics we obtain from measurements of single samples probably represent the larger populations from which the samples were drawn. Almost every statistic has a *standard error*. A standard error is an index number that leads us to conclusions about how far the statistic derived from the sample probably differs from the value we would obtain if we had measured an entire population. A *population* is a well-defined group of individuals or of observations. For example, it could be one composed of Wistar-Institute albino rats between the ages of 30 and 60 days. Or it could be all possible reproductions a certain observer could make of a line 10 centimeters long under the same conditions of rest, time of day, and method of reproduction, *e.g.*, by drawing a line with a pencil. A sample in either case would be a limited

number of observations out of the entire population. Arriving at conclusions that can be generalized to all members of a population depends upon reducing discrepancies between population values and sample values to as small size as possible. This is probably best illustrated by the public-opinion polling, in which the margin of error of voting outcome can be expressed in terms of a percentage of error.

In connection with sampling statistics, there is much in this volume on testing hypotheses. Scientific investigation proceeds from hypothesis to hypothesis. There are numerous hypotheses but relatively few established facts of a general nature. The sooner the research student realizes this point, the better for his clear thinking. There are some investigators, many of them well experienced, unfortunately, who do not make this distinction between a hypothesis and a fact; they mistake hypotheses for facts. For example, there is the hypothesis, stemming from Freudian psychology, that children suffering from asthma are of the "oral-dependent" type and that the breathing spasms are expressions of a cry for aid and for love. The plausibility of the idea, and its apparent consistency with other ideas, may be sufficient to lead many a clinical or psychiatric investigator to act as if the problem were solved; as if the idea were a fact. The properly skeptical investigator makes a study of a sample of asthmatic children and of their nonasthmatic siblings to see whether there is any greater incidence of dependency among the one group than among the other. Probably the most fruitful scientific investigations, at least those that lead to dependable answers, or those that go beyond the exploratory stages, start by setting up a hypothesis, or several alternative hypotheses. Conditions are then arranged in such a way that if the results turn out one way, the hypothesis, or one of its alternatives, is supported and other hypotheses are rendered doubtful. The results must usually be cast in a statistical form which makes possible a decision between hypotheses.

The simplest example of this is seen where we are studying the effects of one thing on another. Let us suppose that it is the effect of benzedrine on ability to reason. We restrict our problem to two alternative and mutually exclusive hypotheses: (1) that benzedrine will affect thinking output or efficiency and (2) that it will not. The first hypothesis can be subdivided into two; that thinking will be facilitated and that thinking will be hindered. The typical experimental operations would be somewhat as follows, briefly described. We develop or adapt a test of reasoning power. We select two groups of individuals of comparable age, education, and *IQ*, both of the same sex. We determine that they are equal on a preliminary trial of the reasoning test. We administer the

drug to one group and a control dose, or placebo, to the other. Neither group knows which has taken the drug. We administer another form of the reasoning test. We obtain two average scores and there is some difference in a certain direction. The question is, does this obtained difference support hypothesis (1) or hypothesis (2)? Could the difference have occurred by chance? If not, it must have been due to the drug, for so far as we know there is no other difference between the two groups that could account for it. It requires a test of the statistical significance of the difference to permit us to reject one hypothesis and accept the other. Having rejected the idea that the difference was due to chance, we may accept the idea that it was due to the drug. Without the statistical test we would be rather helpless in reaching a dependable answer.

The Normal Distribution Curve.—Every student is familiar with the normal distribution curve; it is ubiquitous in psychological and educational literature. There has been much use and abuse of it, and many erroneous things are said about it. The curve itself is a mathematical conception; it does not occur in nature; it is not a biological or a psychological curve. It is an ideal pattern which we can *apply* to useful purpose in many a situation. The distinction between statistics and applied statistics (like that between mathematics and applied mathematics) must be kept in mind. Many fruitful applications of the normal distribution curve in psychology and education will be described in later chapters. These applications are usually made without proof that human variations are normally distributed but with the assumption that they are normally distributed in order that we may benefit from the use of the mathematical properties of the normal curve. If there were knowledge about distributions of human qualities to the contrary, we would, of course, forego these applications. Familiarity with the normal curve and its properties is therefore essential. Its chief applications are seen particularly in Chs. 12 and 19.

Prediction and Statistics.—Three chapters are organized under the heading of "prediction." Most textbooks of beginning psychology start out by saying that it is the purpose of psychology to predict and control human behavior. From that point on, not much more is said about prediction. Dealing with the very complex and intricate set of phenomena that behavior of living organisms presents, and realizing the limitations to accurate predictions, it is appropriate for us to be modest on the subject. We should not feel guilty, however, about our failures to make predictions comparable with those in the physical sciences to the extent that we repress candid and realistic efforts to achieve the

predictions that are possible, nor should we disparage our accomplishments in that direction.

The operation called *prediction* is actually made even when we do not realize it. The vocational counselor who tells a client that he should consider seriously vocations *P*, *Q*, and *R* and should shy away from vocations *V*, *U*, and *W* is tacitly predicting success in the one group and failure in the other. The clinician who diagnoses a person as having an anxiety neurosis is saying that he expects of this individual certain behavior. If he prescribes a certain program of therapy, he is predicting improvement under that treatment versus lack of improvement if it is not applied. The promotion of a child to the next higher grade is a prediction that he will probably adjust better to that assignment than to reassignment to the same grade. Thus, almost all therapies and administrative decisions are, in effect, predictions, whether those who make those prescriptions would be willing to put themselves on record as making predictions or not.

All predictions in psychology and education are what we often call *actuarial*. That is, they are made on a statistical basis and with the knowledge that only "in the long run" will the practice that each prediction stands for be better than otherwise. Prediction of the single case is recognized as being involved with many chance elements. For the single case, the prediction is correct or it is incorrect, depending upon standards. In predicting in large numbers, there are certain probabilities of being right and being wrong. The degree of rightness or wrongness can then be determined. Statistical methods provide the basis for choosing what prediction to make and also a basis for knowing what the odds are for being right or wrong. The various ways of making predictions and the ways of determining their degree of accuracy will be treated at great length in Chs. 14 to 16.

Test Practice and Statistics.—Because tests play such an important role in psychology and education, considerable attention has been given to them in this volume. Recent thinking by statistical psychologists and educators has almost revolutionized our former understanding of tests as instruments of measurement. We may expect in the next ten years considerable progress in working out the implications of these suggestions. Many of the findings have been reflected in the chapters treating tests, particularly Chs. 17 and 18. Certain ideas of reliability and validity of tests had become rather securely intrenched in the thought and practice of test users. These ideas are reexamined and the newer experiences have been used to advantage in the applications of statistics to test practice.

The Student's Aims in His Study of Statistics.—With this overview of content and with the preceding view of the needs and advantages of sta-

tistics, what should the student, particularly the beginner, aim to do about it? In the opinion of the author, the beginner's aims may be listed as follows, in order to make his task more specific.

1. *To master the vocabulary of statistics.*—In order to read and understand a foreign language, there is always the necessity of building up an adequate vocabulary. To the beginner, statistics should be regarded as a foreign language, which he should resolve will not for long remain entirely foreign. The vocabulary consists of concepts that are symbolized by words and by letter symbols that are substituted for them. Along with mathematics in general, statistics shares the ordinary symbols for numerical operations. Thus, much of the vocabulary is already known to the student. As for the new concepts, their meaning will continue to grow the more the student uses them.

2. *To acquire, or to revive, and to extend skill in computation.*—Although it was stated earlier that it is not an important aim for the student to become a statistical clerk, computation is important. For many people, the understanding of the concepts themselves comes largely through applying them in computing operations. The mere step-by-step activities with numbers, when certain goals are in mind, provide opportunities for new insights to occur. The average investigator is never free from a certain amount of computation work to be done. Computation skill, and this includes application of formulas as well as planning efficient operations, like any skill, grows with practice. If there is discouragement at first, further attempts should correct that.

3. *To learn to interpret statistical results correctly.*—Statistical results can be useful only to the extent that they are correctly interpreted. With full and proper interpretations extracted from data, statistical results are a most powerful source of meaning and significance. Inadequately interpreted, they may represent wasted effort. Erroneously accounted for, they are worse than useless. It is the latter eventuality that leads to the common sour-grapish remark "Anything can be proved by statistics." In the hands of skilled operators, statistics make data "talk." It is therefore very important that the implications of any statistical result be realized and that their proper meaning be made manifest. The average reader is less able to interpret the result than the investigator should be. Upon his shoulders rests the responsibility of telling the reader what the conclusions should be and to include, also, some indication of the limitations of those conclusions.

4. *To grasp the logic of statistics.*—Statistics provides a way of thinking as well as a vocabulary and a language. It is a logical system, like all mathematics, which is peculiarly adaptable to the handling of rational

problems in science. This is hard to explain to the beginner. It is hoped that it may become more apparent as later chapters, particularly those dealing with sampling errors, hypotheses, predictions, and factor analysis, are encountered. The most efficient investigator is the one who masters the logical aspects of his research problem before he takes recourse to experiment or to field study. Proper formulation of a research problem is more than half the battle. Too many inexperienced investigators think of a question or a problem and rush to gather data before knowing what it is they really want to observe. Because it is realized that data of some kind must be collected, much time and effort are wasted in collecting the data, without thinking through the problem and coming to the proper decision as to just what kind of data are needed. Or, data are collected in such a manner that no statistical operations now known are adequate to treat the data so as to extract an answer. *Well-planned investigations always include in their design clear considerations of the specific statistical operations to be employed.*

5. *To learn where to apply statistics and where not to.*—While all statistical devices have their power to illuminate data, each has it limitations. In this respect the average student will probably suffer most from lack of mathematical background, whether he realizes it or not. Every statistic is developed as a purely mathematical idea. As such, it rests upon certain assumptions. If those assumptions are true of the particular data with which we have to deal, the statistic may be appropriately applied. The student should note wherever a new statistic is introduced that there are likely to be mentioned certain assumptions or properties of the situation in which that statistic may be utilized. Unfortunately, one can encounter masses of numbers that look as if they are candidates for the use of a certain statistic, *e.g.*, a biserial coefficient of correlation (see Ch. 13), when actually to apply the statistic would be meaningless if not actually misleading. The student without mathematical background will have to learn these exceptions by rote memory or be satisfied with common-sense reasons. He probably would prefer to avoid making ridiculous applications, and when in doubt he should seek advice or refrain from the doubtful application.

6. *To understand the underlying mathematics of statistics.*—This objective will not apply to all students. But it should apply to more than those with unusual previous mathematical training. Many an intelligent student who has not been introduced to analytical geometry or calculus can nevertheless grasp many of the mathematical relationships underlying statistics. This will give him more than common-sense understandings of what goes on in the use of formulas. For the student with mathe-

matical background and for all others who wish to know more about the underlying basis of statistics encountered in the following chapters, the best single source is to be found in the book by Peters and Van Voorhis listed below. We cannot take space to duplicate such proofs in this volume. There are provided in the Appendix, however, a few mathematical derivations of formulas. The selection has been controlled by two considerations: (1) the only mathematics required to follow the proofs is that of ordinary algebra and basic calculus, and (2) the proofs are not readily available elsewhere, either because they do not appear elsewhere or because the sources are scattered.

Some Suggested Aids in Learning Statistics

Following are a few practical suggestions to support the material in this volume.

A Review of Arithmetic and Elementary Algebra.—Some students who have not kept alive the skills they once acquired in arithmetic and elementary algebra frequently feel the need of aids in reviewing those subjects, short of the employment of tutors. To such a student it is strongly recommended that he consult H. M. Walker's *Mathematics Essential for Elementary Statistics*, New York: Holt, 1934. This little volume provides an excellent review in the form of selected exercises of the things that are most needed and in which many students show forgetting. The book is especially recommended to the student who has forgotten his high-school algebra.

Statistical Workbooks.—For the first and second semesters' courses in which this text is used, the student will find useful the two volumes by J. P. Guilford and C. Lovell, *Elementary Statistical Exercises*, Beverly Hills: Sheridan Supply Co., 1946, and *Advanced Statistical Exercises*, by the same publisher, 1950. The first accompanies Chs. 2 through 8 and part of 9. The second covers much of the remaining material of this volume.

Computational Aids.—The wise student will make as much use as possible of all available mechanical aids in the form of calculating machines, tables, and the like. There are inexpensive slide rules now available that will serve when three-place accuracy is sufficient, and this will take care of a large part of one's computations. *Barlow's Tables*, New York: Spon and Chamberlain, are admirable for supplying squares, square roots, and reciprocals for numbers from 1 to 12,500. J. W. Dunlap and A. K. Kurtz have provided many charts, tables, and formulas in their *Handbook of Statistical Nomographs, Tables, and Formulas*, Yonkers-on-Hudson, N.Y.: World, 1932. Where great accuracy in numerical values based upon the normal curve is desired, the recommendation is the monograph by T. L. Kelley, *The Kelley Statistical Tables*, New York: Macmillan, 1938.

Other Books on Statistics.—Other statistical books in which the student or investigator in psychology and education will find coordinate and supplementary reading are as follows:

EDWARDS, A. L. Statistical analysis. New York: Rinehart, 1944.

EZEKIEL, M. Methods of correlation analysis. 2d ed. New York: Wiley, 1941.

FISHER, R. A. Statistical methods for research workers. 8th ed. Edinburgh: Oliver & Boyd, 1941.

GARRETT, H. E. Statistics in psychology and education. 3d ed. New York: Longmans, 1947.

Goulden, C. H. Methods of statistical analysis. New York: Wiley, 1939.
Guilford, J. P. Psychometric methods. New York: McGraw-Hill, 1936.
Holzinger, K. J. Statistical methods for students in education. New York: Ginn, 1928.
Kelley, T. L. Statistical method. New York: Macmillan, 1938.
———. Fundamentals of statistics. Cambridge: Harvard University Press, 1947.
Lindquist, E. F. A first course in statistics. Boston: Houghton Mifflin, 1938.
———. Statistical analysis in educational research. Boston: Houghton Mifflin, 1940.
Peters, C. C. and Van Voorhis, W. R. Statistical procedures and their mathematical bases. New York: McGraw-Hill, 1940.
Snedecor, G. W. Statistical method. 3d ed. Ames, Iowa: Collegiate, 1940.
Treloar, A. E. Elements of statistical reasoning. New York: Wiley, 1939.
Walker, H. M. Elementary statistical methods. New York: Holt, 1943.

CHAPTER 2

COUNTING AND MEASURING

Two Kinds of Numerical Data.—Numerical data generally fall into two major kinds. Things are counted and this yields *frequencies*, or things are measured and this yields *metric values*, or *scale values*. Data of the first kind are often called *enumeration data* and data of the second kind are called measurements or *metric data*.

Statistical procedures deal with both kinds of data, which is the reason for this chapter. There are certain fundamental ideas about numbers and their use that it is well to have in mind before we go ahead. Perhaps it may seem strange to the reader, who has been counting and measuring as long as he can remember, that we should have to devote an entire chapter to these topics. The experts, who, we will have to admit, have had a great deal more experience with numbers and their use than most of us have had, never cease to report new ideas and insights as to the properties of the number system and as to its applications. It is well to keep in mind, incidentally, that there is a real difference between the number system, as such, and its application to counting and measuring. Much confused thinking has resulted from ignoring this fact. The world does not necessarily owe its existence to number and quantity. Numbers were invented by man as a symbolic system of internally consistent ideas which he can use effectively in describing the world as he knows it, thus gaining control over it.

Data and Statistics.—Before we go further, there are some frequently used terms that should be defined. These words are *statistics* and *data*. The word *statistics* itself has several meanings. On the one hand it stands for a branch of mathematics which specializes in enumeration data and their relation to metric data. That is the meaning in the title of this book.

Another meaning, popular but not used by technical people, is implied in the mother's statement when she says, "Bobbie, stay out of the street, or you will become a vital statistic." Here the term in the singular refers to a fact of classification, which is a chief source of all statistics. What the mother meant is that Bobbie would change classification from the category "living" to the category "dead." The keepers of vital statistics in the department of health and in other governmental agencies

would have one less case among the living and one more among the not-living. This use of the term "statistics" is more common among those agencies that keep the records. The numerical records *are* the statistics. While this use of the term is recognized by teachers and writers who specialize in statistics as a subject, their use of the term and the use of it in this book will usually mean something else. In the textbook and classroom situation, we are more inclined to use the word *data* in referring to details in the numerical records or reports. The fact that Bobbie is classified either among the living or the not-living is a *datum*. The word *data* always refers to more than one fact.

In the textbook and classroom situation, too, the singular term *statistic* is most likely to mean a derived numerical value such as an average, a coefficient of correlation, or some other single descriptive concept. It may refer either to the *idea* of an average, a median, a standard deviation, etc., or to a particular value computed from a set of data. The reader can usually tell from the context which usage of these terms is meant.

Data in Categories

Probably most social data are in the form of categorical frequencies; the number of cases in defined classes or categories. The number of births, marriages, and deaths constitutes the bulk of the so-called vital statistics. The number of accidents, fatal or otherwise; the number of arrests for different reasons; and the number of new cases of poliomyelitis constitute other important information by which social agencies keep a finger on the pulse of human affairs. Political and economic interests also have their "barometers" for keeping informed of the trend of events, though some of these depend upon measurements of variables as well as upon counting cases.

Classification.—Before we count, in order to accumulate useful information, we must know what it is we count. We do not count indiscriminately. The frequency that we record refers to a particular class of objects, and this involves the process of classification. Classification of objects has been going on since Aristotle and even before Aristotle. It is a basic psychological process which can be seen in rudimentary form even in the simplest conditioned response. Wherever discriminations are made, along with generalizations, classification of a sort occurs. Useful classifications for counting purposes, however, depend upon a high type of logical analysis. Much of science, following Aristotle, has been of the classificatory type. The classification of plant and animal life into species, genus, and order is the best example. Things thus become ordered and principles emerge.

As science progresses, it is likely to abstract *variables* from its data; continuous variations in single directions. This provides the way for more and more refined measurements. In spite of this general trend in a science, however, the classification of phenomena will probably never cease to be useful. Besides, there are some absolute categories that seem not reducible to continuous variables—life and death; married and unmarried; male and female; and voter and nonvoter. Such discrete classes must be recognized and are usefully dealt with in research as well as in public affairs. Classification, then, is a very useful and necessary process in science as well as in practical life. It is the procedure by which objects become categorized for counting.

Some Psychological Categories.—Before specifying the way in which categories should be set up and utilized, it may be well to have in mind some examples of the more common kinds from the field of psychology. In experimental psychology, particularly in psychophysical studies, we have categories of judgment. The second of a pair of stimuli is judged as "greater than," "equal to," or "less than" the first. In public-opinion polling, responses are obtained in a small number of categories that are intended to be meaningful for interpretation purposes. In answer to the question, "Are you in favor of the Marshall plan?" the response might be "Yes," "No," "I do not know what the Marshall plan is," or "I know what the plan is but I am undecided." In taking a vocational-interest test the examinee may be required to respond in one of three categories, "L" (for like), "I" (for indifferent), or "D" (for dislike) concerning the thing proposed. In a problem-solving experiment with rats, after some preliminary observations, solutions might be categorized as falling into one of four types. Clinical types in psychopathology are categories mostly of long-standing recognition. And so one could continue. Many categories used in research are not static; they change as new light is thrown on the field of study. Some categories are invented for temporary duty as provisional scaffolding upon which to arrange data for better inspection.

There is not space here to give detailed instructions on how to choose or to construct useful categories.[1] It may suffice to say, and it may seem trite to do so, that categories should be *well defined*, *mutually exclusive* (if possible), *univocal*, and *exhaustive*. The importance of good definitions cannot be overestimated. Making proper assignment of cases to classes depends upon it. Being understood by one's colleagues also depends upon it. A prime requirement of scientific findings is that they shall be

[1] For further details on this subject, see Peatman, J. G. Descriptive and sampling statistics. New York: Harper, 1947. Ch. 2.

communicable to others. Other investigators should be able, if they so desire, to repeat our operations to test our results. The requirement of mutual exclusiveness is perhaps the most difficult to achieve. Lack of it probably means something is missing in defining the basis of classification. Lack of it means some overlapping, interdependence, and loss of power to draw clear-cut conclusions. A set of unique categories means that there is one and only one basis of classification. To group school children into three classes, boys, girls, and Mexicans, is to inject two principles or bases: sex difference and race difference. Perhaps anything as grossly absurd is easily avoided; it is the more subtle confusion of variables that causes trouble. By being exhaustive, a set of categories provides a place for all cases. If there are only two classes, such as delinquents and nondelinquents, and if they are well differentiated by objective criteria, even two categories can be exhaustive. In many a system, particularly when more than two classes are needed, there is often a necessity for one miscellaneous group. This group is distinguished merely on the basis of failure to place its members anywhere else. These cases are often ignored, but if they are numerous it probably means biased sampling in other categories. It also probably means lack of adequacy for the classificatory system as a whole.

Qualitative and Quantitative Categories.—Most of the examples of categories given thus far have been what we call *qualitative*. The classes of objects are different in kind. There is no reason for saying that one is greater or less, higher or lower, better or worse than another. The basis is some qualitative attribute. There may be some intrinsic or some external basis for thinking of the classes as being ordered on a scale of more or less, but if so, we are unaware of it. There are, however, many classifications in which the groups can be ordered according to quantity or amount. It may be that the cases vary continuously along a continuum that we recognize but on which we cannot yet make measurements for lack of an instrument; we can only group in a gross manner. Ratings on a scale of five points (and even more) may well be regarded as such a categorizing. In such situations, the categories cannot be defined, perhaps, in any independent terms. Each one may be distinguishable merely by the fact that similar groups of cases are in it and these differ notably from members of other classes.

Another instance is where the experimental controls are in graded steps. Five groups of subjects receive different amounts of instruction of a certain kind. In selection by means of tests, examinees are categorized into the accepted and the rejected groups. Later, after training or service on the job, there is a further classification between those who are satisfactory

and those who are not. Experimental and technological practice is full of such examples. Later chapters will explain methods for dealing with them. The very next chapter will show how metric data are most conveniently handled by somewhat arbitrary groupings in successive categories.

Frequencies, Percentages, Proportions, Ratios.—A *frequency* has already been defined as the number of objects in a category. There are some other related concepts that, though common in advanced arithmetic, most students do not appreciate fully. They play an important role throughout this volume. We cannot review all the arithmetical features of these concepts here, but there are certain new uses of them that should be stressed and certain pitfalls to be pointed out.

Let us consider an example to illustrate the use of percentages. In Table 2.1 are given some original data in the form of frequencies in 12

TABLE 2.1.—ELIMINATION RATES FOR BOMBARDIER STUDENTS OF THREE LEVELS OF APTITUDE IN FOUR ARMY AIR FORCES TRAINING SCHOOLS*

School	Aptitude level											
	Low			Moderate			High			All levels		
	Number in training	Number eliminated	Per cent eliminated	Number in training	Number eliminated	Per cent eliminated	Number in training	Number eliminated	Per cent eliminated	Number in training	Number eliminated	Per cent eliminated
A	62	26	41.9	340	105	30.9	162	29	17.9	564	160	28.4
B	69	23	33.3	274	51	18.6	125	10	8.0	468	84	17.9
C	69	20	29.0	334	43	12.9	166	15	9.0	569	78	13.7
D	139	21	15.1	274	19	6.9	149	9	6.0	562	49	8.7
All schools..	339	90	26.5	1,222	218	17.8	602	63	10.5	2,163	371	17.2

* Aptitude was measured in terms of a composite score on psychological tests. The data were selected from results during the early months of World War II. (Adapted from unpublished data of the AAF Training Command. This will be true of other AAF data used in this volume unless otherwise specified.)

categories. The categories are in a two-way classification, one qualitative and the other quantitative. The data pertain to the number of students in training and the number of these eliminated in each of four bombardier schools in the Army Air Forces during the early part of World War II. In each school the students had been categorized in three levels as to aptitude. The categorization by schools is qualitative and that by aptitude is quantitative. Such a table would probably be set up to study the relation of elimination rate to aptitude and also to differences between

schools. We can make comparisons both ways. There will be some comments, a little later, on how to prepare a good table. Here we are interested in another point: the use of percentages.

Percentage as a Rate Index.—If we wanted to compare schools as to eliminations, the *number* eliminated in each school would be a poor index, particularly when our comparison is made at somewhat constant levels of aptitude. For example, at the low level of aptitude, the numbers of eliminations were not very different: 26, 23, 20, and 21. If we gave credence to such small differences, we should place the schools in the rank order *A*, *B*, *D*, and *C*, from most to least eliminations. Schools *A*, *B*, and *C* had comparable numbers in training, but school *D* had about twice as many. This makes us suspicious of the use of mere numbers eliminated as the way to compare schools. To put the schools on a fair basis we need to find an index of elimination *rate*. We should ask what the elimination "scores" would have been if all schools had had equal numbers in training. If we assume that common number in training to be 100, the number eliminated per hundred is a familiar percentage. The percentages of eliminations for students of low aptitude are 41.9, 33.3, 29.0, and 15.1. Twenty-six is 41.9 per cent of 62; 23 is 33.3 per cent of 69; and so on. Now we see that there are larger differences (this is partly because three of the denominators, 62, 69, and 69, are less than 100) between schools and the rank order is now *A*, *B*, *C*, and *D*. The inversion of the order of *C* and *D* is decisive; at least *D*'s position below *C* now seems decisive. The point of this illustration is that percentages are used to compare groups of objects on an equitable basis. Frequencies alone will not do when such comparisons are to be made.

Some Limitations to the Use of Percentages.—Some precautions should be pointed out concerning the use of percentages. Ideally, a percentage of any number less than about 100 should be computed with hesitation. If the number is less than 100, a change, by chance, of only one case added to or removed from a category would mean a change of more than one per cent. If we ask what per cent 15 is of 25, the answer is 60. But if the frequency were to gain one, the percentage would be 64. If a lower limit must be mentioned as a total below which computation of percentages is unwise, it might be placed at 20. At this number, a change of one case would mean a corresponding change of 5 per cent. This is being quite liberal for the sake of applying a very useful index.

In line with the discussion above, it would seem to be not very meaningful to report percentages to any decimal places unless the total number of cases exceeds 100. When we want a percentage for use in further computations, however, it would be wise to retain at least one decimal

place. Frequencies are "exact" numbers (see p. 33), and percentages based upon them are accurate to as many decimal places as we wish to use. They thus describe the sample in terms of *per hundred*. It is when we become interested in letting an obtained percentage stand for a population value (see Ch. 9) that we must become conservative about reporting it. In Table 2.1 all percentages were reported to one decimal place because most of them were based upon totals greater than 100 and all were made consistent. Consistency of this sort carries some weight, but should not be pushed too far.

When a percentage turns out to be less than 1.0 (*e.g.*, .2 per cent), it is not so meaningful as larger ones, and what is worse, it may be mistaken for a proportion (all proportions are less than, if not equal to, 1.0). In some social statistics a series of percentages may be this small. In this case it is common practice to change the base from 100 to 1,000 or even more, *e.g.*, to report 15 deaths per 100,000; 5 cases in a thousand; and the like. As percentages these would read .0015 and .5, respectively. To avoid confusion with proportions, these should be written as 0.0015 per cent and 0.5 per cent.

Proportions.—Whereas with percentages the common base is 100, with proportions the base, or total, is 1.0. A proportion is a part, or fraction, of 1.0. A proportion is 1/100 of a percentage, and a percentage is 100 times a proportion. Careless individuals often call a percentage a proportion and vice versa. By definition, and in all strictness, they are different concepts. The symbol used for percentage is capital P; for proportion the symbol is a lower-case p. This should help to fix the idea of the relative sizes of the two. The *proportion* of eliminees among low-aptitude students at school A was .419 (see Table 2.1); for high-aptitude students at school B the proportion of eliminees was .080.

As compared with percentages, proportions have some advantages as well as disadvantages. They are less familiar to nonmathematical individuals than are percentages. Whenever results are reported to the general reader, then, percentages are almost always to be preferred. Percentages have another advantage in that we can speak of percentage of gain or of loss. Proportions are always parts of something and can never exceed the total, which is 1.0. They have no place in expressing gain or loss, though presumably losses could be expressed in terms of proportions if we chose, for losses cannot exceed the total; but we never use a proportion for this purpose.

The advantages of proportions are best seen in later chapters. They are used more than percentages, in connection with the normal distribution curve, in connection with item analysis of tests, and with certain

correlation methods, and so on. It has already been said that percentages may be mistaken for proportions when they are less than 1.0. Since proportions can never be greater than 1.0, they are much less likely to be mistaken for percentages.

Probabilities.—Another advantage of proportions is their relation to *probabilities*. Every probability can be expressed in the form of a proportion. We say that the probability of getting a head in tossing a coin is 1/2 or 1 chance in 2. This is a more manageable figure if expressed as a probability of .5. We say that in throwing a die the probability of getting a six spot is 1 in 6. Expressed as a proportion this is .167. In general, for computation purposes, decimal fractions are much preferred to common fractions; they are much more easily manipulated in addition and subtraction and in finding squares and square roots. The interchangeability of proportions and probabilities will be found to be a very common occurrence in the later chapters.

Ratios.—A ratio is a fraction. The ratio of a to b is the fraction a/b. A proportion is a special ratio; the ratio of a part to a total. We may also have ratios of one part to another. For example, there were 69 low-aptitude students in training school B (Table 2.1), of whom 23 were eliminated and 46 were graduated. The ratio of graduates to eliminees was 46/23, or 2 to 1. This ratio can also be expressed as 2.0. The ratio of eliminees to graduates was 23/46, or .5. This could also be expressed as .5 to 1, but ordinarily is not. At any rate, in a ratio the base is 1.0, as it is in a proportion. The chief difference is that a proportion is restricted to the ratio of part to total, whereas ratios are not.

Ratios are useful as *index numbers*. They describe rates and relationships. The *IQ* is an index number of rate of general mental growth—the ratio of mental age to chronological age (multiplied by 100). Comparisons of incomes of regions are made in terms of per capita—the ratio of total income to population. Costs of education are more meaningful if stated in terms of dollars per pupil per day attended rather than in terms of total sums of expenditures. In dealing with index numbers one should keep in mind the operations by which they were derived. It sometimes makes a difference when they are used in computation as in averaging them or in correlation problems (see pp. 355 and 358).

Tabulation of Data.—Every student who writes a report based upon data is faced with the problem of how best to organize them in tables. Tables serve several purposes. There are tables that list the raw or original data. Lists of scores in several tests earned by different individuals provide an example. Although these may be very long in some reports, many readers like to see them presented in full so that they may

apply checks or perform other operations than the investigator used. One common way to present these tables is in an appendix to the report.

A second type of table is a summarizing device. It is used to present an organized and curtailed picture of what is in the original data. It includes such descriptive statistics as means, standard deviations, and the like, with the data grouped in one or more meaningful ways. Table 2.1 is an example of this type. All the essential information is there. Such a table should tell a complete story of its kind. It should be given a title that tells clearly what the table is about. If the title becomes too long it is better to relegate to a footnote some of the secondary information. Headings of columns and rows should be descriptive, and their spacing and the lining should show clearly to what columns or rows they belong. A table should be so labeled that the reader need not turn to the text material in order to know what is there.

How to Prepare Tables.—The organization of such a table, in columns and rows, should take into consideration, first, what are the main points that should be brought out. In Table 2.1 probably the more important comparison to be made is that of the different schools. A person concerned with the administrative aspects of bombardier training certainly would think so. One who is concerned with the development of aptitude tests would, of course, be interested in the other problem, the relation of elimination rate to aptitude level. In the latter case, a distinction between schools would be of little importance. Having decided which relationship is of most interest, the data should be arranged so that the comparisons one wants to make most are easiest to observe. Here let us say we want to compare schools. The best basis of comparison is in terms of elimination rates. The four elimination rates are in an uninterrupted column. Comparison of elimination rates for different aptitude levels is more difficult because other numbers intervene.

A second consideration, and it is of less importance, is the practical one of keeping the dimensions of the table consistent with the dimensions of a page. Columns can be longer than rows; consequently, considering the space available for headings and the widths of numbers, we can fit the data in the available space. With small tables this is no problem. Ordinarily, long lists go better in columns and short lists in rows. Another consideration is the psychological fact that horizontal eye movements are easier and more natural for a reader than are vertical movements. All these considerations must be weighed and balanced against one another.

A third type of table is a final, summarizing one. This brings together the salient findings from several tables. The second type may, of course, serve the same function; it all depends upon the scope and nature of the

study. If there is a final-type table, however, it serves as a basis for major conclusions of the study.

Graphic Representation of Data.—The graphic representation of data has become such an extensive art that it is possible to provide only an introduction to the subject here. A few fundamental principles will be mentioned and illustrated. A "picture may be worth ten thousand words" but only if it is properly done. The first requirement is that it shall tell a complete story for what it is intended to convey.

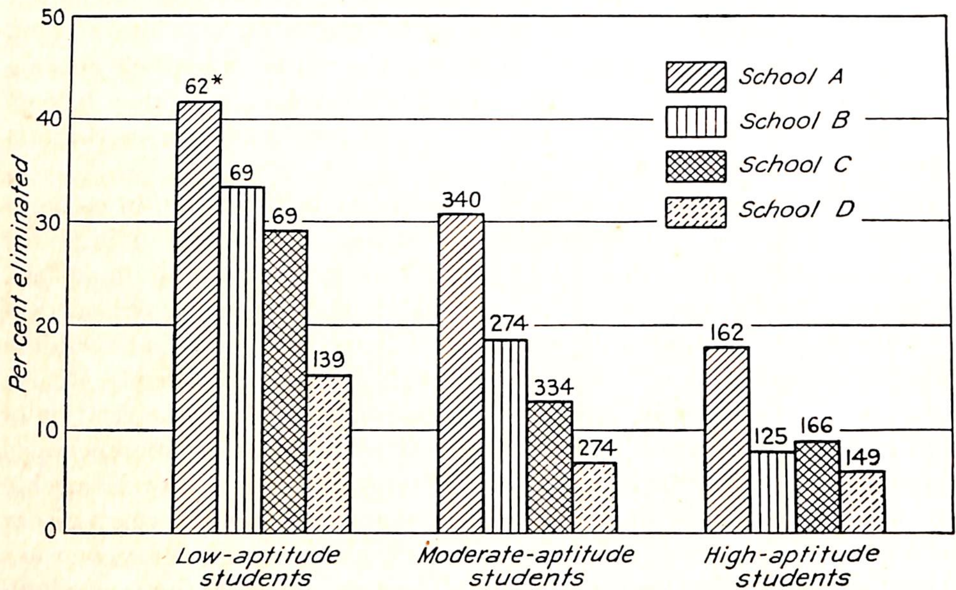


FIG. 2.1.—Percentage of bombardier students eliminated from training in four different Army Air Force schools during the early part of World War II. Comparisons are made at three different aptitude levels.

Bar Diagrams.—Probably the most common type of figure for displaying frequencies or percentages for categories is the *bar diagram*. It is very adaptable to many purposes and arrangements.

Figures 2.1 and 2.2 are designed to represent the data of Table 2.1. In these examples, the bars are in the vertical position, but bars can also be placed in the horizontal position (Figs. 2.3 and 2.4). In Fig. 2.1 the data are grouped so as to show best a comparison of the different schools. There are three groups of bars, one for each level of aptitude of students, and within each group every school is represented. In each case, the same kind of shading is used for the same school. The schools were arranged, in general, in their order of elimination rate. They should be

in the same order in the three groups. This facilitates cross comparisons between aptitude levels and gives an idea of trend within each group. Figure 2.2 was designed to emphasize comparison of elimination rates as dependent upon aptitude level. There are four groups, one for each school, with three bars in each group. Here the quantitative nature of the aptitude variable determines the order of the three bars in each group.

In both diagrams, note that the numbers of students in training are given at the tops of the bars. The statistically minded reader will want

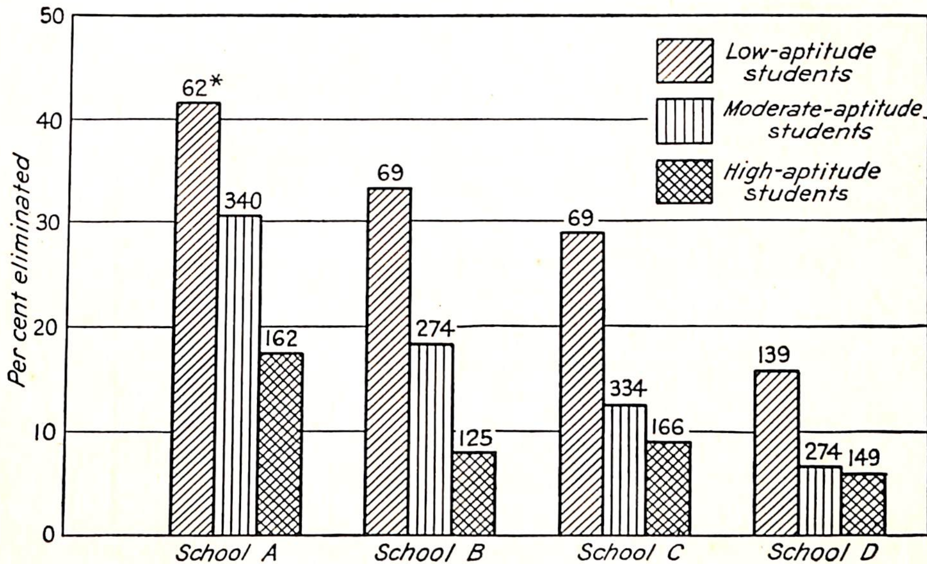


FIG. 2.2.—Percentage of bombardier students eliminated from training at three different levels of aptitude. Comparisons are made in four different bombardier schools of the Army Air Forces during the early part of World War II.

to know these values as a basis for judging about how reliable each percentage is and whether differences he sees in the bars are probably genuine or are perhaps due to chance. He cannot be sure about these questions unless he applies some procedures described in Ch. 9, but he can get a rough idea just by knowing the total numbers and by general past experience.

Figures 2.3 and 2.4 show some data in response to the question, "How many times did you feel afraid while flying on a combat mission?" applied to aircrew personnel just returning from combat to redistribution stations in the United States.[1] The categories of responses were "Every time,

[1] From the publication, Wickert, F. (Ed.). Psychological research on problems of redistribution, AAF Aviation Psychology Program Research Reports, No. 14, 1947.

or almost every time," "About ¼ to ¾ of the times," "One to three times," and "Never." This is not the place to question either the method or the validity of the responses. We are merely illustrating a statistical device. In Fig. 2.3 the bars are designed to compare officer with enlisted aircrew personnel. For each category of response the bars for these two kinds of personnel are shown juxtaposed. The numerical percentage values are also written in so that the reader will have the more accurate information that numbers provide if he wants it. The sizes of samples

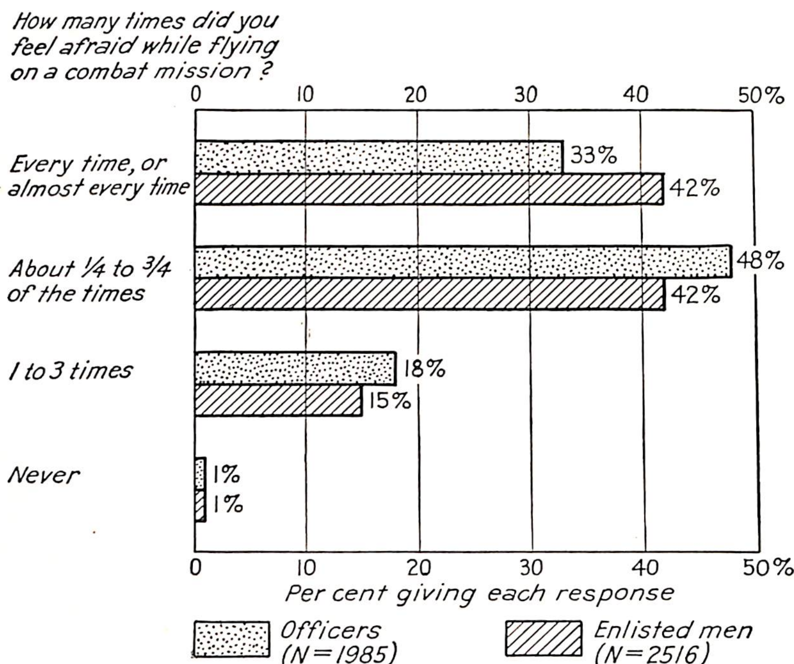


FIG. 2.3.—Percentages of officer versus enlisted personnel in samples of Army Air Forces combat returnees who responded in specified ways to a question concerning fear in combat.

are given below the diagram so that the reader may have some basis for degree of confidence in the differences represented.

Figure 2.4 shows another arrangement of the same data. In this diagram we obtain a better conception of the proportions of reactions in each category for officers as a group and for enlisted men as a group, as well as some possibility of comparing the two in each category because the two bars are presented parallel and the category percentages in the same rank order.

Pie Diagrams.—Another kind of picture that is sometimes used to show proportions of a total is the *pie diagram*. The 360 degrees of a circle are subdivided in proportion to the number or percentage in each category.

Figure 2.5 is an illustration. It shows the situation with regard to aviation cadets in the AAF with respect to three principles of classification: previous flying experience, marital status, and training preference. The number in the total sample is given below each diagram. The numerical percentage is written in each segment which is shaded differently from

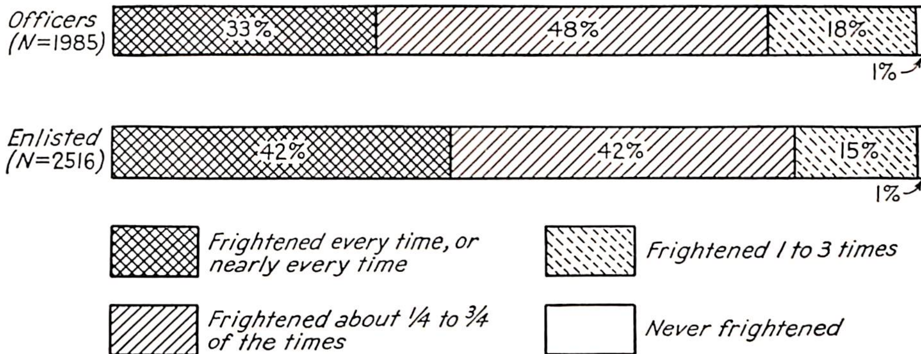


FIG. 2.4.—Percentages of responses of each type given to the question, "How many times did you feel afraid while flying on a combat mission?" by samples of officer and enlisted personnel who had returned from tours of combat duty in the Army Air Forces.

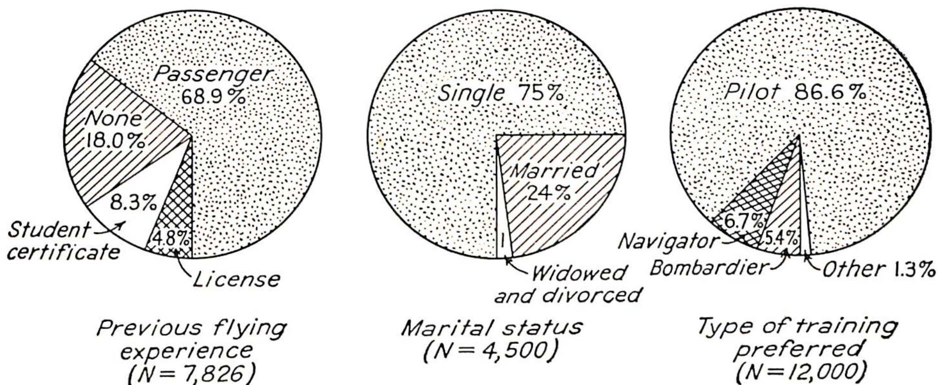


FIG. 2.5.—Descriptions of the status of new recruits to flying training in the Army Air Forces during the early part of World War II with respect to previous flying experience, marital condition, and type of training preferred.

others in the same "pie." The category name is also written in a segment if there is room; if not, it is written just outside.

The pie diagram is restricted to this kind of display; the proportions of a total. It is inferior to the bar diagram, such as that in Fig. 2.4 (which also demonstrates proportions of wholes), when we want to compare the same categories in two samples.

Trend Charts.—When showing changes in frequencies, percentages, or proportions over a period of time, a *trend chart* or *belt graph* is desirable.

One could show a bar for each sample and place the bars in time order, but this would not picture changing conditions nearly as well as something continuous. Fig. 2.6 is drawn to represent such changing conditions or trends in a certain situation. The data are in terms of percentages of aviation students interviewed, who were subsequently recommended to different types of assignment. The data arose from the psychological unit at one classification center during World War II and cover a period of 15 months during the last part of 1942 and the first part of 1943.

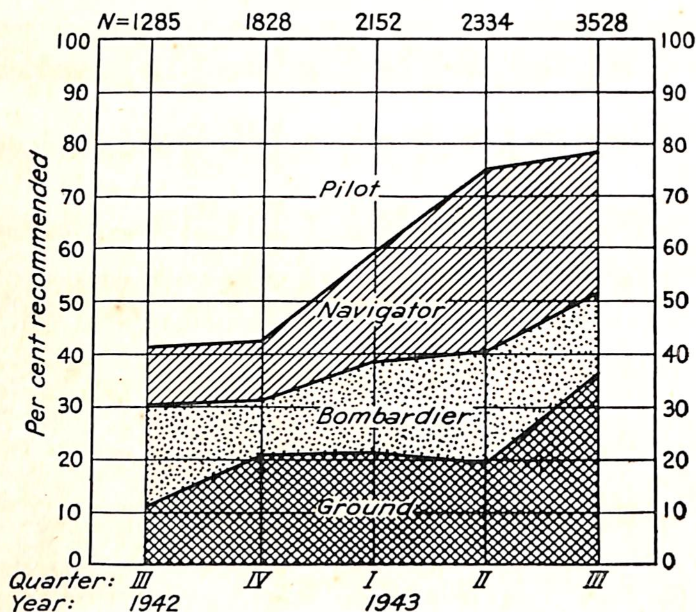


Fig. 2.6.—Trend in the percentages of interviewed aviation students in the Army Air Forces who were recommended for various assignments during a fifteen-month period of World War II. (*Adapted from data in the AAF Aviation Psychology Research Program, Report No. 2, The classification program, P. H. DuBois, Ed.* P. 346.)

Observations were grouped by quarters, or three-month periods. The students interviewed were those whose classification on the basis of aptitude scores and expressed preferences for different types of training was not obvious under the prevailing regulations at the time.

In some trend charts the *frequencies* are plotted—for example, those representing population growths or those representing changes in income. In connection with the data of Fig. 2.6, we are not interested in numbers but, for administrative reasons, in proportions of students disposed of in each of four ways, for assignment to one of three types of training or to ground duty. The reasons for any trends are, of course, not obvious from the picture itself, but knowing the picture, a study of the situation would

probably yield an explanation of the causes and suggest, if necessary, corrective measures.

There are other trend charts of various kinds. In a broad sense, all curves of learning and retention would be included. Their nature is so well known that they need not be described here.

Pictographs.—The layman, who is probably not interested in statistics or numbers, can be induced to read reports and to gain impressions the

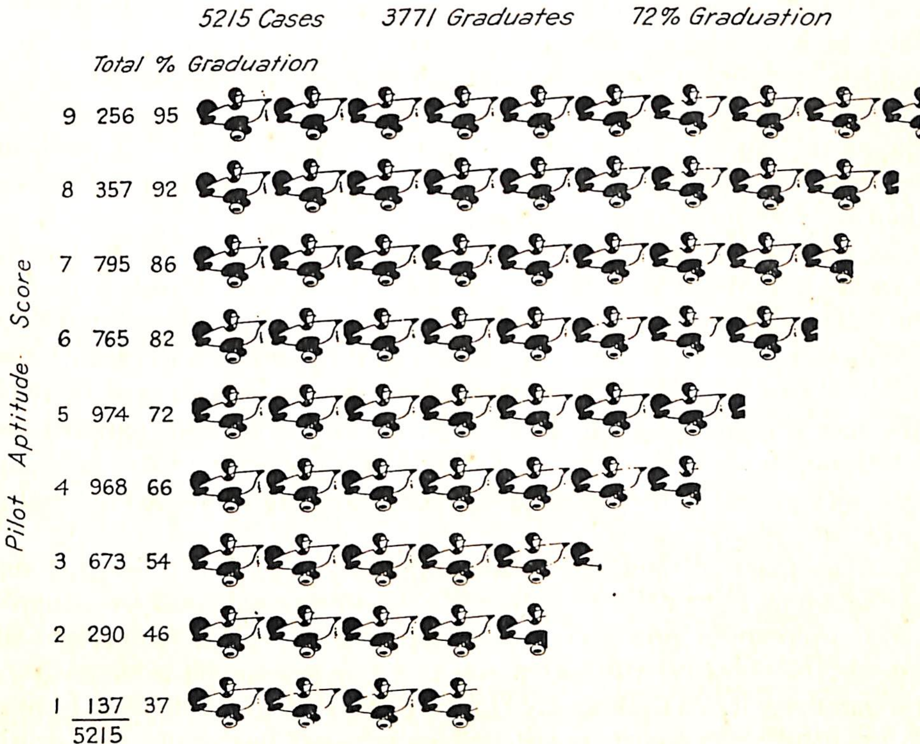


FIG. 2.7.—Percentage of pilot students at each aptitude level who graduated from primary training in one sample of Army Air Forces trainees during World War II. (*From Aircrew selection and training, a publication of the AAF Training Command Headquarters*, 1944.)

writer wishes to make, if the picture is dressed up in terms of concrete objects. Figure 2.7 is one example that was used to display to the average reader the relationship that existed at the time between graduation rate and aptitude of pilot students in the AAF. It requires a minimum of statistical sophistication to interpret such a picture, and the cartoonish quality of the drawings attracts attention and interest. Very effective reporting of statistical results to the general public is done in this manner. The number and variety of ways in which this can be applied are limited only by the ingenuity of the reporter.

Measurements

Some Examples of Psychological Measurements.—In order to make our discussion concrete and specific, let us consider some typical examples of measurements commonly made by psychologists. Perhaps the first examples that come to mind are scores on tests of mental ability. These are usually in terms of the number of correct responses to test items. A similar kind of measurement is seen in scores on a personality questionnaire or a vocational-interest inventory. In these cases it is not the number of "correct" responses but the number of responses indicating the same interest or trait, often weighted in proportion to their supposed diagnostic value. Also in the area of mental tests we find the frequent reference to "chronological age," "mental age," and that ratio between the two, the "intelligence quotient."

In the experimental laboratory as well as in the clinic, we frequently measure in terms of the time required to complete a specified test or task. In memory experiments, we measure learning efficiency in terms of the number of trials to attain a certain standard of performance or in terms of the "goodness" of performance at the end of a certain trial or time. We measure efficiency of retention in terms of the time required for relearning (overcoming the forgetting that has taken place) and the efficiency of recall in terms of association time or in terms of the number of items correctly recited.

In the sphere of motivation, we gauge the strength of drive in terms of the amount of punishment (electric shock) an organism (for example, a rat) will endure in order to reach his immediate goal or in terms of the number of times he will take a constant punishment in order to attain the same result. The difficulty of a task or test item can now be specified in quantitative terms, as can the affective value (degree of liking or disliking) for a color, a sound, or a pictorial design. In studies of sensory and perceptual powers, the threshold stimulus and the differential limen are given in terms of stimulus magnitudes. The span of perception or of apprehension is given in terms of the average number of items that the observer can report correctly after momentary exposures. The galvanic skin response, the pupillary response, and the amount of salivation also serve as quantitative indicators of amounts of psychological happenings.

Some Examples of Educational Measurement.—Many an educational problem is also a psychological problem, and its mode of measurement has been indicated in the preceding paragraphs. Achievement in any area of learning, like any mental ability, is measurable in terms of test scores. Marks, however obtained, have been the traditional mode of evaluating

students in specific units of formal education. Attendance records, data on size of classes, on budgets, on supplies, and on other material aspects of the well-regulated school system compose another list of measurements in education. Outcomes of educational effort are often expressed quantitatively in terms of promotion statistics, achievement ratios, and estimates of teaching success. Whether for purposes of research in education or for systematic and meaningful record keeping, statistical methods become indispensable tools.

Some Different Kinds of Measurement.—In a superficial way, it is easy to see, as one glances over the list of psychological and educational measurements just mentioned, that there are different kinds of measurement involved. Among the psychologist's measurements, some are in terms of the stimulus—for example, the threshold stimulus or stimulus difference; the number of syllables or items; the amount of electric shock, etc. Others are in terms of the amount of response—for example, time of the response; number of responses or of correct responses; degree of the response, etc. Some measurements are more direct, such as reaction time, and others more indirect, such as affective value and difficulty. Some measurements are in terms of discrete units—number of individuals, syllables, words, items, crossings—and others are in terms of continuous scales—age, time of response, amount of punishment, and degree of effort. In the discrete type of measurement, things can increase or decrease only by changing one whole unit at a time, whereas in the continuous type, the increase or decrease can be by as small a fraction of a unit as one pleases and can distinguish. Although this difference has a logical significance, in statistical practice, actually, we generally treat discrete and continuous measurements in the same manner.

Rank Orders and Other Measurements.—In a most general sense, we make a measurement whenever we assign numbers to things in such a way that those things are placed in order. Suppose we place three boys, Charles, Bob, and David, in rank order for height, Charles, being rank 3 (tallest) and David, rank 1 (shortest). The numbers 3, 2, and 1, attached to Charles, Bob, and David, give us some useful information, such as the inference that Charles is taller than Bob and that Bob is taller than David. These numbers do not tell us much more. Since they are merely ranks, we cannot say that Charles is as much taller than Bob as Bob is taller than David. We cannot say that Bob is two times as tall as David or that Charles is three times as tall as David. Measurements in terms of rank order simply give us the serial arrangement of things.

As we saw from the example just given, we are not at liberty to add and subtract or to multiply and divide such numbers. Had we actually

applied a meter stick to these three boys and found that their heights were: Charles, 195 cm., Bob, 180 cm., and David, 150 cm., matters would be different. Now we can make some further deductions about the heights of these boys. We can say that the difference between Bob and David is two times that between Bob and Charles. Knowing that Charles is 15 cm. taller than Bob and that Bob is 30 cm. taller than David, we can infer that Charles is 45 cm. taller than David. We can say that Bob is 20 per cent taller than David and that Charles is 30 per cent taller than David. It is apparent that we can now perform all the arithmetical operations of addition, subtraction, multiplication, and division with the three numbers assigned to the three boys.

Best Measurements Require an Equal Unit and an Absolute Zero.—Some measurements obtained in psychology and education are comparable with the measurements of height (linear distance) just mentioned, but most are not. Many measurements should be regarded as merely placing things in rank order until it is demonstrated that they give us more accurate information than that. We have something considerably better than rank order when our measuring scale possesses equal units. When this is true, a gain of a unit in one part of the scale is equal to a gain of a unit in any other part of the scale. We can then perform a number of different operations with numbers assigned to objects on such a scale that would otherwise be precluded.

A measuring scale is not complete, however, unless it also has an absolute zero point. An example of a scale that has equal units but not an absolute zero point is the centigrade thermometer. The zero point is arbitrarily placed at the freezing point of water. With this instrument, we can say that the temperature of the weather changes as much when it rises from 0 to 25 as it does when it rises from 25 to 50. But we cannot say that 50° is twice as warm as 25° or that 100° is twice as hot as 50°. We can find differences between numbers on this scale and get sensible answers, but we cannot multiply and divide. If we translate our zero mark to the absolute zero point (zero heat), which in terms of the common thermometer is −273°, then we can perform these operations. On the absolute scale, our 25° becomes 298°, and our 50° becomes 323°. Now it is obvious that the higher of the two (323) is not two times the lower (298). But if our absolute centigrade scale is correct, with regard to equality of units, we may well say that a temperature twice as hot physically as 298° is a temperature of 596° (also on the absolute scale).

Mental-test Scales as Metric Devices.—What shall we say of a measuring scale of the type most frequently used in psychology and education—mental-test scores in terms of number of items correct? Have

we here a scale with absolute zero and equal units? Strictly speaking, usually not. A score of zero, no items correctly answered, does not mean zero ability. For had we included some easier items, even the lowest individual in the test could probably have made a score numerically greater than zero. Thus we are unable to say that a score of 50 points means twice the ability represented by a score of 25 or half the ability represented by a score of 100 points. For if our real zero-ability score should have been some 25 points below our arbitrary one, these three scores would then become 50, 75, and 125.

Now the second is *not* twice the first or half the third. Nor can we be sure that our units are equal within the range of scores obtained. Unless the units were equal, we should not be able to say that a score of 100 is as far above one of 75 as the latter is above a score of 50. As a matter of long experience, however, we find that test scores generally behave as if units were equal; as if one item correct adds an amount to the measurement of ability equal to that added by any other item correct. There are various indications that tell the experienced worker in statistics when his measurements probably possess equal units and when they do not. And when they do, we can proceed to apply most of the ordinary statistical procedures. When we strongly suspect that they do not, we can make adjustments or substitute other statistical methods that do apply. The beginner in statistical work need not be too much concerned about trying to decide the matter, but he should be aware that there are natural limitations to what one may do in the way of statistics and that most of our ordinary conclusions are sound only in so far as equal units (and much less often an absolute zero point) prevail in the measuring scale.

How Numbers Should Be Regarded in Measurement.—Most measurements are taken to the nearest unit—nearest foot, inch, centimeter, or millimeter, depending upon the fineness of the measuring instrument and the accuracy we demand for the purposes at hand. In giving the height of a tree, measurement to the nearest foot—for example, 107 ft.—would be adequate. In giving the height of a girl, we should resort to inches or perhaps centimeters as our practical unit. In giving the length of a needle, we should probably report in terms of millimeters; and in giving its diameter as seen under a micrometer, we should resort to some smaller unit. In any case, we may notice that our object does not contain an exact number of our chosen units. Our tree is more than 107 ft. but is closer to 107 than it is to 108; our girl is not exactly 156 cm. but is closer to 156 than to 155; etc. The result is that our report of 107 for the tree means anything between 106.5 and 107.5 ft., and our report for the girl

means anything between 155.5 and 156.5 cm. Figure 2.8 shows a graphic illustration of units and their limits.

And so it is with most psychological and educational measurements. A test score of 48 is taken to mean from 47.5 to 48.5; and an obtained score of 70 means from 69.5 to 70.5. We assume that a score is never a point on the scale but occupies an interval from a half unit below to a half unit above the given number. We can make this seem more reasonable by arguing that the person making a score of 48 actually might be just a fraction of a unit better than 47.5 at the moment, and being better than 47.5 is sufficient to give him a whole score of 48. Or our individual might just fail to be as good as 48.5 on the same test, but, not being quite good enough to achieve 49 items, he falls back to 48. Although our tests are probably never so refined as to cause an individual to waver between fractions of a point (the margin of error is usually more than a

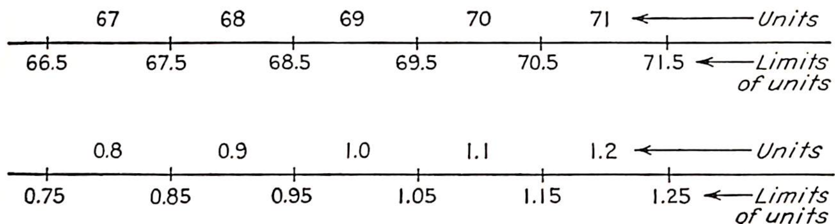


FIG. 2.8.—An illustration of two metric scales, showing selected units and their limits.

whole point), this kind of argument rationalizes our procedure from one standpoint.

A more important practical consideration dictates *the taking of a score as occupying a whole interval on the scale*, as the student will appreciate later. If we did not do this, an average computed from a set of ungrouped measurements would not be consistent with one computed when the same measurements are grouped. Even in dealing with discrete measurements, as, for example, the number of children in a family, we customarily proceed *as if* 8 children meant anywhere from 7.5 to 8.5. The only notable exception to this general rule is in dealing with chronological age as given to the *last* birthday and the like. Then a twelve-year-old child is anywhere from 12.0 to 13.0. If ages are given *to the nearest birthday*, however, our rule again applies, and a twelve-year-old falls in the interval 11.5 to 12.5.

Some Rules Regarding Numbers

Approximate and Exact Numbers.—Measurements, when taken to the nearest whole unit, are known as *approximate numbers*. They are always

"fuzzy" and are of uncertain value within the unit where they fall. When we find a number by enumeration of discrete objects, we have an exact number; for example, 15 men, 42 letters, or 50 pencils. The distinction between exact and approximate numbers we shall find important when they are used in calculations. Some rules about calculations are presented next. They would be unnecessary if all numbers in statistics were exact.

How to Round Numbers.—The beginner in statistical computation invariably asks, "How many decimal places shall I save?" In just this form, the question cannot be answered. The question should read instead, "How much accuracy have I in the answer?" A number may have been rounded, dropping *all* digits to the right of the decimal point, yet not all of the remaining figures may be accurate. Another number may have four places remaining to the right of the decimal point, yet all of them may be accurate. Some students may, if they lack good rules, drop too many figures, thus losing much of the accuracy that they really have; others may save a string of figures beyond the limit of accuracy, giving the appearance of great exactness that is really fictitious.

First let us be clear as to the proper way to round a number. There is no particular difficulty in rounding to the nearest whole number; 15.7 becomes 16, and 27.4 becomes 27; 9.6 becomes 10, and 0.96 becomes 1. In rounding to two decimal places, the same principles apply; 2.1827 becomes 2.18, and 91.2179 becomes 91.22. It is when the first digit to be dropped is 5 that difficulties arise. In rounding to two decimal places, again, the number 7.1654 becomes 7.17, and even 7.16502 becomes 7.17 rather than 7.16, for the reason that the decimal fraction beyond the 6 is greater than just .00500. Had the number been 7.16499, we should have rounded to 7.16, because it is a shade closer to 7.16 than to 7.17.

When the number is 7.16500 (equidistant between 7.16 and 7.17) we follow an arbitrary rule that when the digit preceding the 5 is an even number we leave it as it is but when this number is odd we raise it to the next digit. Thus 7.16500 would be rounded to 7.16, but 7.17500 is rounded to 7.18. The main reason for this is that when such numbers are summed, in a long series, we should have had by chance as many that were raised a half point as were lowered the same amount, and the changes will tend to compensate for one another.

A word should be added about leaving a rounded number ending in the digit 5. For example, the number 6.21499 rounded to three decimal places becomes 6.215. Were we to round this further, following our rule, we should have 6.22. In view of the original number, this would be incorrect. It would have been well to indicate when the number 6.215 was given

that the 5 came by rounding upward or that the original number was less than 5 in the third decimal place. We can do this by writing it as 6.215 — to show this fact. The number 42.5+ has been rounded from something greater than 42.50. Further rounding to a whole number gives 43, in spite of the odd-even rule offered above.

How Many Significant Figures in a Number?—When a measurement is given as 107 ft., the number is not only accurate to the nearest unit but is also said to be accurate to three significant figures. In spite of the fact that this measurement was taken only to the nearest foot, the 7 fixes the value between 106.5 and 107.5, which makes the 7 significant. If we had, instead, a measurement of 107.3 ft., there would be accuracy to the nearest tenth of a foot and four significant figures. The .3 added to the number now fixes the measurement between 107.25 and 107.35 ft., tying the last place to the .3 ft.

The number .00156 has just three significant figures or digits. They are the only ones that tell us about the numerical value, the two zeros being required merely to locate the position of the decimal point. The number 15600, likewise, has only three significant digits, again the two zeros merely being used as "fillers" to locate the decimal point. If this were given as the approximate cost of a certain boat in dollars, we should conclude that the cost was anywhere from 15550 to 15650 dollars. But if it had been written as 15600., with a decimal point after the last zero, this would indicate that measurement was to the nearest unit, or within the limits of 15599.5 to 15600.5 dollars.

When zeros come between other digits, they count as significant figures. Thus 1002.1 has five significant figures, and .071021 also has five. Any other zero not used to fix the decimal point is also usually significant, as in .420, which has three significant digits, since the last digit fixes the number between .4195 and .4205. A lone zero before the decimal point, however, as 0.41, is not significant, since it adds nothing to our information concerning numerical value.

Rules Governing Significant Figures in Computation.—The following rules will determine how many significant figures there are in a number found by computation.

1. *In Sums of Numbers.* CASE I.—When all the numbers added are regarded as accurate to the nearest unit, the sum is regarded as accurate to the nearest unit.

Example: 47 + 161 + 5,171 = 5,379, a sum that is accurate to the nearest unit and that has four significant figures.

A similar case occurs when all the numbers added have the same number of decimal places.

Example: 2.91 + 40.22 + 0.07 = 43.20, where the answer is accurate to the second decimal place because all the numbers were accurate to that place.

CASE II.—When numbers that are not accurate to the same number of places at the right of the decimal point are added, the sum is accurate only as far as the number having the *smallest* number of decimal places.

Example: 17.257 + 142.1 + 75.47 = 234.8, which is rounded from 234.827. Note that the rounding was done *after* summing and not before.

A similar rule is true when numbers rounded to the *left* of the decimal point are summed.

Example: 75,000 + 3,845 = 79,000, which is rounded from 78,845 because in the first number there are only two significant digits to the left of the hundreds place.

2. *In Differences.* CASE I.—If the two numbers are accurate to the same digit at the right, the difference is also accurate that far to the right.

Example: 173.24 − 98.84 = 78.40, the zero being significant.

Frequently a difference is drastically reduced in the number of significant figures, so much so that further computations with this difference are sometimes lacking in desired accuracy. This situation is to be avoided when possible.

Example: 4.692 − 4.685 = 0.007.

CASE II.—As with addition, the answer is accurate no further to the right than is the number whose accuracy extends less far to the right. In the following examples, the answers are rounded to as many significant figures as are accurate.

Example: 175.1 − 82.715 = 92.4 (not 92.385).
Example: 5,200 − 829 = 4,400 (not 4,371).

In both these cases, contrary to the practice in summing numbers, the rounding can just as well be done before subtracting, for the result will be the same either way.

3. *In Products of Numbers.* CASE I.—The product of two approximate numbers has no more accurate significant digits than has the number with the smaller number of significant digits.

Example: 41.57 × 1.3 = 54 (not 54.041).

CASE II.—The product of an exact number times an approximate number has no more accurate significant figures than has the approximate number.

Example: 24.091 × 22 = 530.00 (where 22 is an exact number).
Example: 24.09 × 72 = 1,734 (where 72 is an exact number).

CASE III.—The product of two exact numbers is accurate to all obtained digits.

Example: 175 × 42 = 7,350 (which may be written as 7,350.).

4. *In Quotients.* CASE I.—The quotient of two approximate numbers has no more accurate significant digits than the one having the smaller number of significant digits.

Example: 7.182 ÷ 2.3 = 3.1 (not 3.12261).
Example: 4.07 ÷ 0.2815 = 14.5 (not 14.458).

CASE II.—The quotient from an exact and an approximate number contains no more accurate significant numbers than the approximate number.

Example: 7.1025 ÷ 22 = 0.32284 (where 22 is an exact number).

CASE III.—The quotient of two exact numbers may be written to as many significant figures as one wishes.

5. *In Squaring a Number.*—Since this is a matter of multiplying a number by itself, the same rules as those governing products will apply. In general, the square of an approximate number contains no more accurate significant figures than the number itself.

6. *In Square Roots of Numbers.* CASE I.—The square root of an approximate number contains roughly the same number of significant figures as the number itself. The square root of 85.7, for example, may be taken appropriately to be 9.26, to three significant figures.

CASE II.—The square root of an exact number may be given to as many places as one wishes.

Example: $\sqrt{5}$ = 2.2361. This could be carried further, or we could round it to 2.236 or to 2.24, depending upon our purposes.

In many statistical problems which the student will encounter, the square root of a number of persons or observations will be utilized (see Ch. 9 particularly). The number of discrete objects is an exact number, thus the square root can be carried as far as one wishes. A good practice to follow is to think how many significant digits are needed for further computation. As a general suggestion, one might use not less than three significant digits in such a square root.

Application of the Rules.—Although the rules as just given are acceptable and sound, one should use them as guides and not follow them slav-

ishly. One frequently has to use his best judgment and do the most reasonable thing. To follow the rules rigidly at every step of the way would sometimes introduce inaccuracies or else cause one to lose information that he really has and needs. One good general principle to follow is to *carry along more significant figures through the successive steps of calculation than would be required for strict accuracy under the rules and withhold the rounding of numbers until the final answer is obtained*, such as an arithmetic mean, a standard deviation, or a correlation coefficient. At the end of a solution, one may decide upon the extent of accuracy in the answer by applying the rules to every step in the series of numerical operations. This is difficult in some problems because of the many steps. There are also other things to be considered in particular situations, such as the standard error (see Ch. 9) of the statistic computed. For these reasons further suggestions will be offered more appropriately later when we are dealing with specific cases.

The student will now see the reason for the earlier statement (p. 33) to the effect that the question "How many decimal places shall I save?" cannot be answered very simply. The most important things to carry away from the discussion above are a better appreciation of the problems of accuracy and, roughly, some of the limitations to accuracy of figures derived from measurements.

Exercises

1. In a certain school in a southwestern city, the fifth grade had 80 pupils, of whom 32 were of white, American-born stock, 20 were of Mexican, 10 of Japanese, and 18 of American-Indian stock. Complete the following table:

Stock	Frequency	Percentage	Proportion
American white............	32		
Mexican..................		25.0	
Japanese..................			.125
American-Indian...........			

2. In the preceding data, what was the ratio of Mexicans to Indians? Of American white to Japanese? Of Indian to American white?

3. In selecting a child at random from the fifth-grade group, what is the probability of getting a Mexican? Of getting a Japanese? An Indian? Either a Mexican *or* an Indian?

4. In the fourth grade of the same school, the following numbers of children appeared: American white, 47; Mexican, 27; Japanese, 11; and Indian 15. In the third grade the numbers were: 66, 30, 6, and 18, respectively. Prepare a tabulation of the data in the three grades. Draw conclusions from the table.

5. Draw bar diagrams representing the racial data given above.

6. Draw a trend chart representing the same data.

7. State the exact limits to the following scores or measurements: 57 sec. 150 kg. 65 score points 0 score points 14.5 cm. .125 sec. 15 years (to the last birthday).

8. Round the following numbers to one decimal place: 26.418 4.072 4.98 9.092 120.052 0.3500 44.7508 291.6500 8.8502 31.15− 48.25+.

9. How many significant figures in each of the following numbers: 1,942 20,007 170.9 0.31 28,000 0.0017 0.3400 21,5000.

10. Write the answers to the following problems to as many significant figures as the rules concerning accuracy allow:

a. 21.3 in. times 15 (where 15 is an exact number).

b. $5.2 + 17.2509 + 918.04$.

c. 242.8×0.075.

d. 4.27505 divided by 25 (where 25 is an exact number).

e. 17.98 divided by 2.1.

f. 38.6 squared.

g. $\sqrt{50}$ (where 50 is an exact number, but be reasonable).

h. $\sqrt{25.3179}$

CHAPTER 3

FREQUENCY DISTRIBUTIONS

After we obtain a set of measurements, the next customary step is to put them in systematic order by grouping them in classes. A set of individual measurements, taken as they come, as in the list in Table 3.1, does not convey much useful information to us. We have merely a vague, general conception of about how large they run numerically but that is about all. The data in Table 3.1 are scores made by 50 students in an ink-blot test. Each score is the number of objects the student reported in observing 10 ink blots during a period of 10 min. Concerning such a set of data we usually want to know several things. One is what kind of score the average or typical student makes; another concerns the amount of variability there is in the group or how large the individual differences are; and a third is something about the shape of the distribution of scores, *i.e.*, whether the students tend to bunch up at either end of the range or at the middle or whether they are about equally scattered over the entire range. The first steps in the direction of answering these questions require the setting up of a frequency distribution.

TABLE 3.1.—SCORES IN AN INK-BLOT TEST

25	33	35	37	55	27	40	33	39	28
34	29	44	36	22	51	29	21	28	29
33	42	15	36	41	20	25	38	47	32
15	27	27	33	46	10	16	34	18	14
46	21	19	26	19	17	24	21	27	16

THE CLASS INTERVAL—ITS LIMITS AND FREQUENCIES

The Size of Class Interval.—We could begin by asking how many scores of 25 there are, of 26, 27, etc., but this would not give us an adequate picture, because in a group of only 50 individuals whose scores range from 10 to 55, many scores do not occur at all and others occur only once. We therefore combine the scores into a relatively small number of *class intervals*, each class interval covering the same range of score units on the scale of measurement.

The first thing to be decided is the size of the class interval. How many units shall it contain? This choice is dictated by two general customs to which experience has led us to agree. *One is the rule that we should have not less than* 10 *nor more than* 20 *class intervals.* Though in rare instances we find workers going outside those limits, the general tendency is for them to keep within the boundaries of 10 to 15. The small number of groups is favored by the fact that we often deal with small numbers of individuals in our measured sample and by the urge for convenience. The larger number is favored by the desire for accuracy of computation, because the process of grouping will introduce minor errors into the calculations, and the coarser the grouping, that is, the smaller the number of classes, the greater is this tendency.

Some Sizes Preferred.—The second rule determining the choice of class interval is that certain ranges of units (scores) are preferred. They are 1, 2, 3, 5, 10, and 20. These six intervals will be found to take care of almost all sets of data. To apply these rules to our data in Table 3.1, we need first to know the total range of scores from highest to lowest. The highest score is 55, and the lowest is 10, which gives us a total range of 46 points (one more than the highest minus the lowest). An interval of 3 points is the one that would give us the best number of classes that our first rule requires. It will be found that the range divided by the number of units in the class interval (in this case 46 divided by 3) ordinarily gives the total number of class intervals needed to cover the range. In this instance, we should therefore have 16 groups. If we chose 5 units as our class interval, we should have 46/5, which is 10 groups. In view of the relatively small number of cases, and because an interval of 5 will give us the minimum of 10 groups, we choose 5 as our class interval.[1]

Where to Start the Class Intervals.—It would be a quite natural tendency to start the intervals with their lowest scores at multiples of the size of the interval; when the interval is 3, to start them with 9, 12, 15, 18, etc.; when the interval is 5, to start with 10, 15, 20, 25, 30, etc. This is by far the most common practice, though it is admittedly arbitrary. When the size of the interval is 3 or 5, there are arguments for starting intervals in such a way that the multiple of the size of interval is in exactly the middle of the group. Thus the grouping by three's would give groups like 8, 9, 10, and 14, 15, 16, etc.; by five's, it would be 8, 9, 10, 11, 12 and 18, 19, 20, 21, 22, etc. The midpoints would be multiples of 3 in the one case and of 5 in the second case. We use score limits so

[1] While the rules as just stated will be satisfactory for most purposes, some variations will be presented later in connection with grouping for graphic representation of distributions and for estimating a mode (see Ch. 4).

much more than we do midpoints, however, that the arguments seem mostly to favor beginning intervals consistently with the multiples of the size of interval, even when the size is 3 or 5 units.

Score Limits of Class Intervals.—We shall follow the usual practice here, placing in the lowest interval all scores of 10, 11, 12, 13, and 14; in the next higher interval, scores of 15, 16, 17, 18, and 19; etc. (see Table 3.2). Instead of writing out all the scores for each interval, we give only the bottom and top scores. Our intervals are then labeled 10 to 14, 15 to 19, 20 to 24, etc., or, more often, 10–14, 15–19, 20–24. The bottom and top scores for each interval represent what we call the *score limits* of the interval. They do not indicate exactly where each interval begins and ends on the scale of measurement. The score limits are useful primarily in tallying and in labeling the intervals.

TABLE 3.2.—FREQUENCY DISTRIBUTION OF THE INK-BLOT SCORES THAT WERE LISTED IN TABLE 3.1

(1) Scores	(2) Tally Marks	(3) Frequencies, f
55–59	/	1
50–54	/	1
45–49	///	3
40–44	////	4
35–39	~~////~~ /	6
30–34	~~////~~ //	7
25–29	~~////~~ ~~////~~ //	12
20–24	~~////~~ /	6
15–19	~~////~~ ///	8
10–14	//	2

$\Sigma f = 50 = N$

Exact Limits of Class Intervals.—We shall soon find that in computations we must think in terms of *exact limits*. Remember that a score of 10 actually means from 9.5 to 10.5, and that a score of 14 actually means from 13.5 to 14.5. This means that the interval containing scores 10 to 14 inclusive actually extends from 9.5 to 14.5 on the measurement scale. Likewise, the interval having score limits of 15 and 19 has exact limits of 14.5 and 19.5 on the scale. The interval labeled 55 to 59 actually extends from 54.5 and 59.5. The same principle holds no matter what the size of interval or where it begins. An interval labeled 14 to 16 includes scores 14, 15, and 16 and extends exactly from 13.5 to 16.5. An interval labeled 70 to 79 extends from 69.5 to 79.5. It will be seen that

by following this principle each interval begins exactly where the one below leaves off, which is as it should be (see Fig. 3.1).[1]

Tallying the Frequencies.—Having decided upon the size of class interval and with what scores to start the intervals, we are ready to list them, as in Table 3.2. It is accepted custom to place the highest measurements at the top of the list and the lowest at the bottom, as shown here. Space is left in the second column for the tallying process. Taking each score in Table 3.1 as we come to it, we locate it within its proper interval and write a tally mark in the row for that interval. Having completed the tallying, we count up the number of tally marks in each row to find the

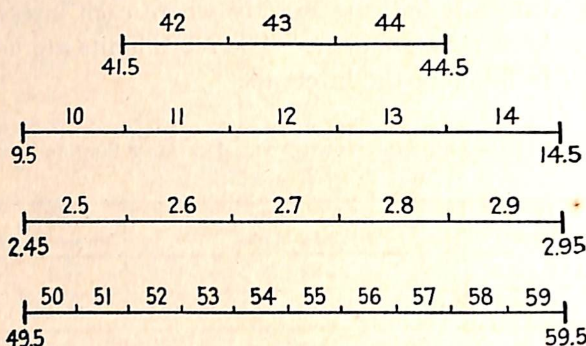


Fig. 3.1.—Exact limits of class intervals with different sizes of interval and of unit of measurement.

frequency (f), or total number of individuals falling within each group. The frequencies are listed in the third column of Table 3.2.

Checking the Tallying.—Next we sum the frequencies, and if our tallying has omitted none and duplicated none, the sum should equal the number of individuals. At the bottom of the column we find the symbol Σf, in which Σ (capital Greek sigma) stands for "the sum of" whatever follows it. Thus, Σf is "the sum of the frequencies." The total number of individuals or measurements in our sample is symbolized by the capital letter N, which stands for "number." If Σf does not equal N, there has been a mistake in tallying, and tallying should be repeated until this check is satisfied. Even if Σf does equal N, there could have been a tally or two placed in the wrong interval. There is no way of checking this kind of error except by doing the tallying twice. The moral is that

[1] Strictly speaking, limits such as 69.5 and 79.5 also stand for very small distances rather than points. Only in a *relative* sense are they division points between intervals. Some writers define an interval such as the one containing scores from 70 to 79 as being actually from 69.5000 to 79.4999. One could extend the zeros and nines indefinitely. For practical purposes the "exact" limits of 69.5 and 79.5 will serve very well when measurements are integers.

great care should be taken to make the finding of frequencies correct at the first attempt.

Graphic Representation of Frequency Distributions

The frequency distribution in Table 3.2, particularly the array of tally marks, gives us a general picture of the group of individuals as a whole. We can see, for example, that the most frequent scores fell in the interval

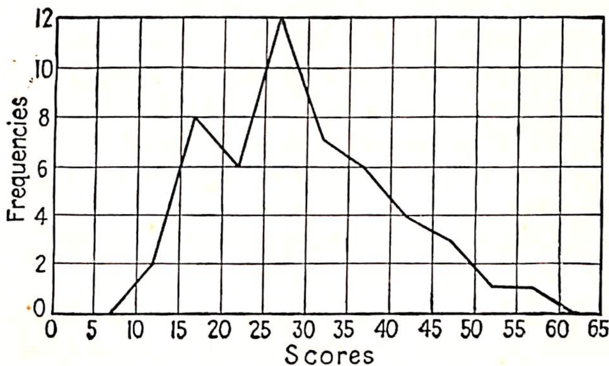


Fig. 3.2.—A frequency polygon for the distribution of scores in the ink-blot test.

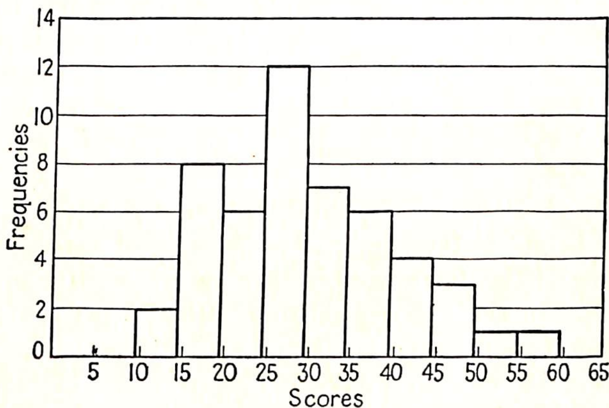


Fig. 3.3.—A histogram for the same distribution as in Fig. 3.2.

25–29, that the very low and very high scores are more rare, and that the greatest bunching of scores comes in the lower half of the range. Much better pictures of this distribution are afforded in Figs. 3.2 and 3.3, however, where the general contour of the distribution is more accurately represented and the numbers of cases in the various intervals are more exactly shown. Fig. 3.2 is of the type known as *frequency polygon*, and Fig. 3.3 is of the type called *histogram*, or sometimes, though less often, *column diagram*.

The Frequency Polygon and How to Plot It.—A polygon is a many-sided figure, and thus the picture in Fig. 3.2 derives its name. There are a number of factors to be kept in mind in drawing such a figure.

The Kind of Graph Paper.—First, it might be said that, in general, the most convenient type of cross-section paper is the type that is ruled into heavy lines 1 in. apart each way, subdivided into tenths of an inch more lightly drawn.

The Width of the Diagram.—Second, the question of the height and width of the entire figure arises. For the sake of easy readability, the width of the figure should be at least 5 in. We have altogether 10 class intervals in which there are frequencies, but in drawing the diagram, we should allow for one more class interval at each end of the scale, making 12 in all. This is to permit bringing the ends of the polygon down to the base line (see Fig. 3.2).

Labeling the Base Line.—In deciding how many intervals to allow to the inch, it is well to remember that we are going to label the base line of the figure in terms of our measuring scale and hence should plan things so that $\frac{1}{10}$ in. will stand for an integral number of units on this original scale. In the ink-blot data, we have been dealing with a class interval of 5 units, and we are making room for 12 intervals on our base line—in other words, for 60 units. By allowing $\frac{1}{10}$ in. to each unit ($\frac{1}{2}$ in. to each class interval), our distribution will spread over an extent of 6 in., which is sufficiently large. On the base line, therefore, we label every fifth line with a multiple of 5, beginning with 5 at the left and ending with 65 at the right.

The Height of the Figure.—The third important question is with regard to the relative height of the figure. For the sake of appearance and also for easy reading of the diagram, there is a general custom of making the maximum height of the distribution from 60 to 75 per cent of the total width. Our total width is 6 in. or $\frac{60}{10}$ in. Sixty per cent of this would be $\frac{36}{10}$ in., and 75 per cent would be $\frac{45}{10}$ in. Our highest frequency, as we see in Table 3.2, is 12. By allowing $\frac{3}{10}$ in. to the person, the height of $\frac{36}{10}$ would be attained, and by allowing $\frac{4}{10}$ in. to a person a height of $\frac{48}{10}$ in. would be reached. The former comes within our rule, and the latter does not; therefore we adopt $\frac{3}{10}$ in. as the unit on the vertical scale.

How to Locate a Midpoint.—In order to plot a dot to represent the frequency in each class interval, we must next decide above what point on the base line the dot shall be. It is plotted exactly at the midpoint of the interval, and the midpoint is exactly midway between the *exact* lower and upper limits of the interval. A simple rule to find the midpoint is to

average either exact or score limits of the interval. The interval containing scores 10 to 14 inclusive has exact limits of 9.5 and 14.5. The entire range is 5 units. Half this range is 2.5 units. Go this far above the lower limit, and you have 9.5 plus 2.5, or 12 exactly, as the midpoint. This could be written as 12.0. Or deduct 2.5 from the upper limit, 14.5 minus 2.5, and you also have exactly 12.0 as the midpoint; or the average of 10 and 14 is 12.0. The midpoint of the interval 55–59 is 57.0. When the class interval is 5 and the lowest score in each interval is a multiple of 5, as will be true in many of the instances met in psychology and education, the midpoints will end in 2 and 7 systematically. For the sake of a complete picture of the midpoints for the data in Table 3.2, we

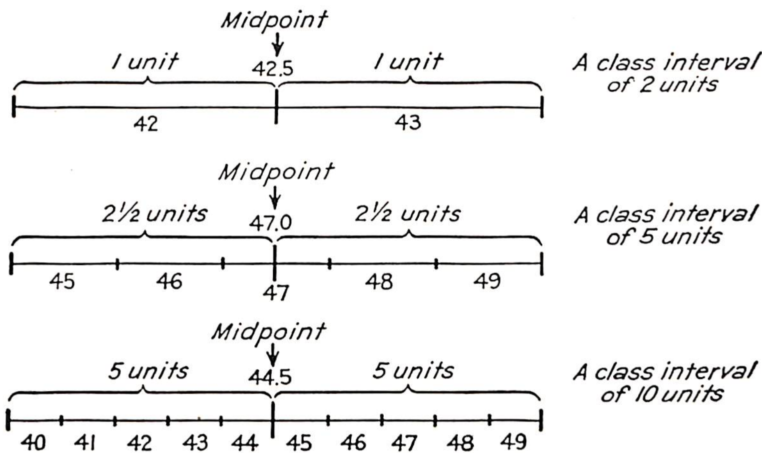


FIG. 3.4.—Midpoints of class intervals with differing numbers of units.

have given in Table 3.3 the full set of midpoints. For a general illustration of midpoints, see Fig. 3.4.

Plotting the Points.—Having determined the midpoints and knowing the frequencies corresponding to them, we are ready to plot the dots for the frequency polygon. For the two intervals at the ends of the distribution (see Table 3.3) we have frequencies of zero. Sometimes there are frequencies of zero *not* in the last two classes. When so, we plot these dots also on the base line and bring the lines that connect the dots down to the base line at those places. That did not happen to be the case in these data. When the dots are placed at the midpoints, as directed, it may be noted that they do not appear directly above the midpoints of the marked places on the base line (5, 10, 15, 20, etc., in this case). Remember that these multiples of 5 are *not* the exact limits of the class intervals; they are merely convenient and meaningful reference points on our original scale. Had we begun the class intervals at scores other than multiples

TABLE 3.3.—CLASS INTERVALS AND THEIR MIDPOINTS

Score limits	Exact limits	Midpoints	Frequencies
60–64	59.5–64.5	62	0
55–59	54.5–59.5	57	1
50–54	49.5–54.5	52	1
45–49	44.5–49.5	47	3
40–44	39.5–44.5	42	4
35–39	34.5–39.5	37	6
30–34	29.5–34.5	32	7
25–29	24.5–29.5	27	12
20–24	19.5–24.5	22	6
15–19	14.5–19.5	17	8
10–14	9.5–14.5	12	2
5– 9	4.5– 9.5	7	0

of 5—for example, at 11, 16, 21, 26, etc.—we should still plot at the midpoints of the intervals (now different than before) and should still label the reference points as multiples of 5, as in Fig. 3.2. The curve as drawn truly represents the shape of the distribution as we have grouped the scores.

The Histogram and How to Plot It.—Many of the facts learned in plotting the frequency polygon also apply in plotting the histogram. The choice of size, proportions, units per square of graph paper all are the same. The only important difference is that although we locate the height of each column or rectangle by placing a dot at the midpoint of each interval, we do not then connect dot to dot with straight diagonal lines. Instead, we draw a short horizontal line through each dot (see Fig. 3.3), extending it to the upper and lower *exact* limits of each class interval. Those exact limits are given in Table 3.3 for our data. Having done this, we erect vertical lines at each of these exact limits tall enough to form complete rectangles. Again it may be noticed that the rectangles seem to be misplaced a half unit with respect to the numbers on the base line, but this is correct; the choice of limits for our classes makes the exact limits come a half unit below the multiples of 5, *i.e.*, at 4.5, 9.5, 14.5, 19.5, etc.

Advantages and Disadvantages of the Two Types of Figure.—On the whole, the frequency polygon seems generally preferred to the histogram. For one thing, it gives a much better conception of the contour of the distribution; the transition from one interval to another is direct and probably describes the distribution more accurately. The histogram

gives a stepwise change from interval to interval, based upon the assumption that the cases falling within each interval are evenly distributed over the interval. The polygon gives the more correct impression that on both sides of the highest point (directly above the mode), the cases within an interval are more frequent on the side nearer the mode, except where there are inversions in the general trend (as between scores of 15 and 25 in Fig. 3.2).

On the other hand, the histogram gives a more readily grasped representation of the number of cases within each class interval; each measurement or individual occupies exactly the same amount of area. One more advantage favoring the polygon is that when we wish to plot two distributions overlapping on the same base line, as, for example, two different age groups or the two sexes, the histogram type gives a very confused picture, whereas the polygon type usually provides a clear comparison.

Plotting Two or More Distributions When *N* Differs.—The comparison of two distributions graphically raises a new question when the numbers of individuals in the two groups differ. With large differences, naturally, there is the question of scale, or how much space to give the figure. If the smaller distribution is large enough to be clearly legible, the larger one may extend beyond reasonable bounds. Furthermore, if it is general shapes and general positions on the measuring scale and dispersions that we wish to compare, the marked difference in size may make such comparisons very unsatisfactory. A common solution to this difficulty is to reduce both distributions to *percentage frequencies* instead of plotting the original frequencies. It is then as if we had two distributions, each of whose *N*'s equal 100. This makes their two areas approximately equal in the polygon form, and comparisons of shape, level, and dispersion are then quite satisfactory.

How to Find Percentage Frequencies.—As an example of how to transform frequencies into percentages the data in Table 3.4 are presented. In each case, the frequencies in the distribution are each multiplied by 100, then divided by *N*. A shorter procedure would be to find the quotient 100/*N* to four or more decimal places, then multiply each frequency in turn by this ratio. In distribution I, the ratio is 100/51, which equals 1.9608, and in distribution II it is 100/160, which equals 0.6250. Multiplying each frequency f_1 by 1.9608, we obtain the list of percentages in column (4), and multiplying each frequency f_2 by 0.625, we obtain the list in column (5). Plotting these percentages above the corresponding midpoints of class intervals, we obtain the distribution curves in Fig. 3.5. Although it was apparent in Table 3.4 that the second group were higher on the scale than the first and that there was still considerable over-

TABLE 3.4.—FREQUENCY DISTRIBUTIONS OF SCORES IN A COLLEGE-APTITUDE TEST FOR FRESHMEN AT TWO DIFFERENT COLLEGES

(1) Scores	(2) f_1	(3) f_2	(4) P_1	(5) P_2
140–149		8		5.0
130–139		32		20.0
120–129		48		30.0
110–119	1	29	2.0	18.1
100–109	0	18	0.0	11.2
90–99	3	14	5.9	8.8
80–89	5	5	9.8	3.1
70–79	6	5	11.8	3.1
60–69	14	0	27.5	0.0
50–59	7	1	13.7	0.6
40–49	11		21.6	
30–39	4		7.8	
Sums.......	51	160	100.1	99.9

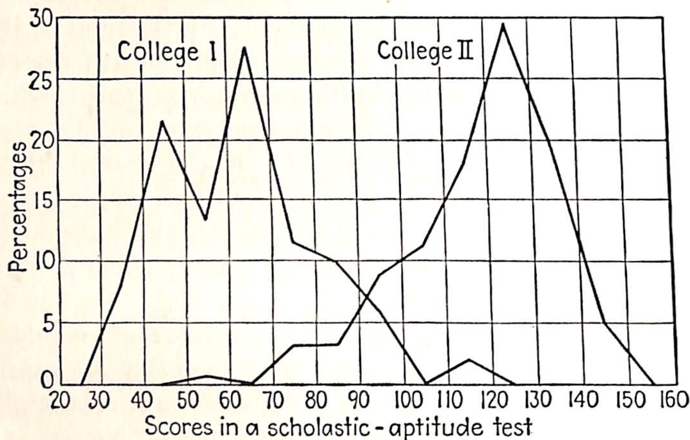


FIG. 3.5.—Distributions of scores in an aptitude test in two colleges. Frequencies have been reduced to a percentage basis.

lapping of scores between the two, these facts are more clearly brought out in graphic form. Also much clearer is the somewhat narrower dispersion in the second group as compared with the first.

Skewed Distributions.—In addition, the fact is more clear that the first group bunches at the left in its own range and has relatively few high scores, whereas the second group bunches at the upper end of its range, with relatively few low scores. We describe the first distribution

as being *positively skewed* (pointed end toward the right or positive direction) and the second distribution as being *negatively skewed* (pointed end toward the left or negative direction). The greater irregularity of contour in the first distribution is probably due to the small number of cases originally in this group. The changing of the two distributions to the percentage basis has not changed the contour, only the general vertical size of the curves.

Comparison of Two Histograms.—The same two distributions as illustrated in Fig. 3.5 may also be shown in the form of histograms. When overlapping histograms become rather involved and confusing, writers

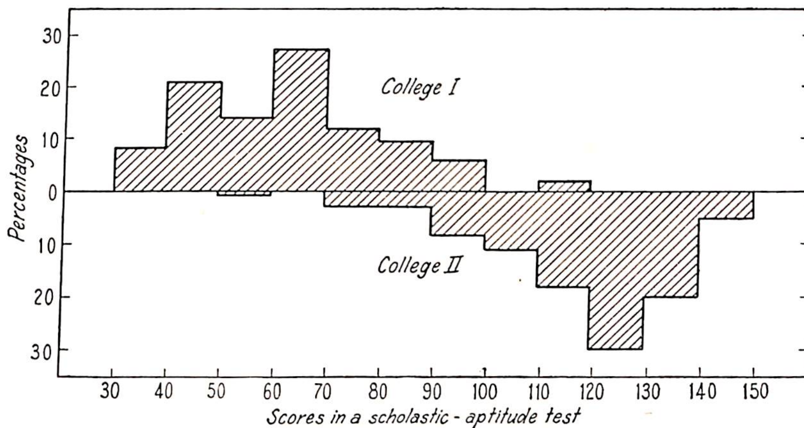


FIG. 3.6.—Same distributions as represented in Fig. 3.5 shown in the form of two histograms.

sometimes resort to the device shown in Fig. 3.6. In that illustration, a mirror reflection is pictured for one of the distributions, but both are drawn on the same horizontal scale. The frequency scale (in terms of percentages here) is repeated, also in mirror reflection. The shading of the rectangles is optional, but it has the virtue of making the entire surface within each histogram stand out from the page.

Other Variations in Presenting Overlapping Curves.—The distributions in Fig. 3.5 are clearly represented as shown in two overlapping polygons. There are certain instances in which such line drawings will not suffice. One of these is when the two distributions are so extensively overlapping that there is considerable crisscrossing of lines and only confusion would result unless something is done about it. Fig. 3.7 demonstrates such a situation and also how the matter is handled, namely, by showing the one polygon in a dotted line. By inspection one can readily

see to which group all parts of a polygon belong. The groups are identified, each with its type of line, by giving the code, in this instance, in the upper right part of the chart. Figure 3.7 also includes desirable information such as is lacking in Fig. 3.5, namely, the total number of individuals in each sample.

Figure 3.8 gives another demonstration of overlapping distributions that call for several different kinds of lines. This is generally desirable when there are more than two polygons on the same chart and when there is any overlapping at all.

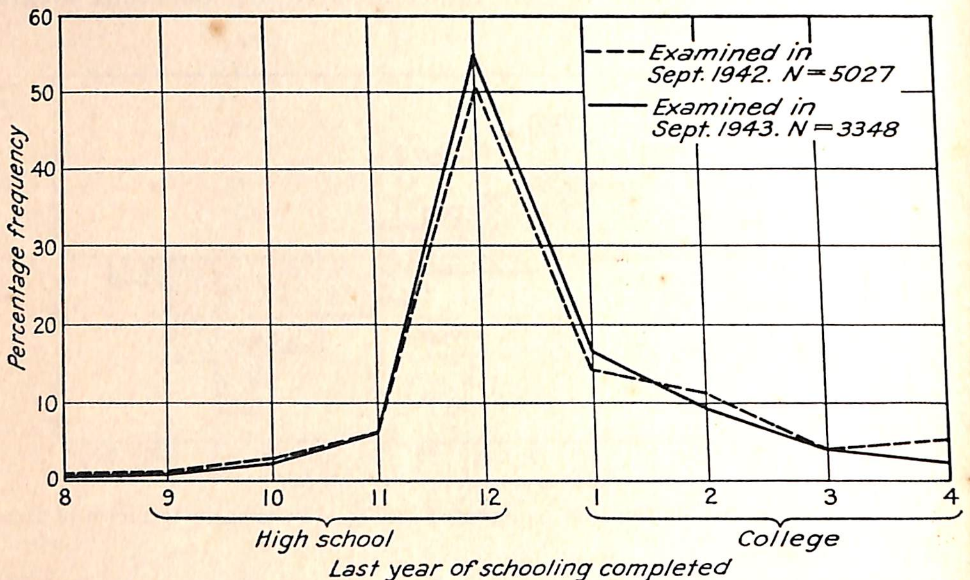


FIG. 3.7.—Two overlapping frequency polygons representing distributions of years of schooling completed by samples of aviation students in the AAF.

Figures 3.7 and 3.8, particularly, demonstrate how much meaning one can extract from pictorial representations of frequency distributions. Questions of policy governing the selection and training of aviation students during World War II hinged upon questions of age and of formal education of recruits, and it was important to maintain a clear picture of changing status of the trainees in these respects. From Fig. 3.7, for example, one would conclude that the typical recruit was a high-school graduate and that men of this category comprised more than half of all recruits. It might have been surprising to some of the commanding officers to find that there were recruits with as little formal schooling as 8 years who could pass the Army Air Forces qualifying examination. Those with less than 12 years of school were in very small percentages,

however, and either this type of man did not apply in large numbers for aircrew training or he was screened out quite generally by the qualifying examination. The fact that the two curves, for samples a year apart, are almost identical throughout indicates that the same kind of men, so far as previous education was concerned, were applying and qualifying for admission to AAF flying training.

The distributions of aircrew recruits as to chronological age (Fig. 3.8) tell quite a different story. Within the same period of a year, although the same range of ages prevailed (it was limited by regulations) there

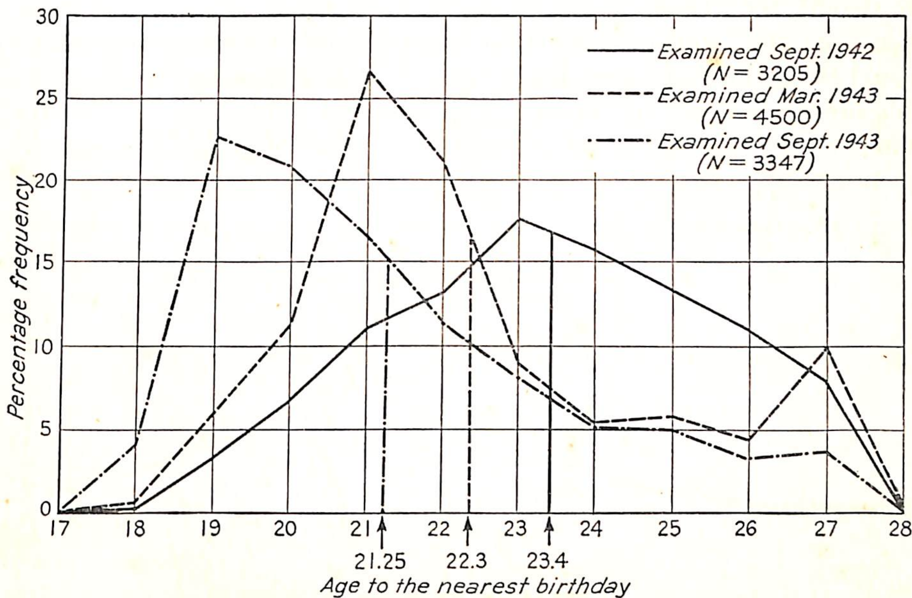


FIG. 3.8.—Three overlapping frequency polygons representing distributions of chronological ages of aviation students in the AAF.

was a drastic trend toward reduction of age. This is shown by the fact that the mode (age having the greatest frequency) was at twenty-three years in the September, 1942, sample, at twenty-one years in the March, 1943, sample, and at nineteen in the September, 1943, sample. The skewing was slightly negative in the earliest sample and markedly positive in the latest sample. In one of the samples there was a secondary mode at twenty-seven years. This reflects the known fact that many twenty-seven-year-old men expedited their entrance into AAF flight training in order to assure acceptance before reaching the age limit.

Smoothing a Frequency-distribution Curve.—Any set of measurements like those in Fig. 3.5 is usually regarded as one sample out of a larger population having practically the same properties as the ones obtained in

the sample. The first group is one of freshmen entering a certain college in a given year. If it is assumed that over a run of years the kind of students seeking entrance and the kind accepted remain about the same, the 51 students whose scores are given here may be said to represent the larger population. Had we obtained similar scores for this larger population, the irregularities seen in Fig. 3.5 would no doubt have been minimized.

We frequently wish to forecast, from the supposed representative sample that we have, how a larger population would distribute itself. To do this, we smooth the frequency distribution in the following manner. We predict from the frequencies we have what the corresponding frequencies would be in the larger population by a system of running averages. In this process, we permit the two frequencies on either side—*i.e.*, in the immediately neighboring intervals—to help determine the expected frequency in any class. In Table 3.5, the obtained frequencies f_o are given

TABLE 3.5.—ORIGINAL AND SMOOTHED FREQUENCIES FOR A DISTRIBUTION OF SCORES IN A SCHOLASTIC-APTITUDE TEST

(1) Scores	(2) f_o	(3) f_e
120–129	0	0.25
110–119	1	0.50
100–109	0	1.00
90– 99	3	2.75
80– 89	5	4.75
70– 79	6	7.75
60– 69	14	10.25
50– 59	7	9.75
40– 49	11	8.25
30– 39	4	4.75
20– 29	0	1.00
Sums.......	51	51.00

in column (2), and it will be noticed that two class intervals have been added at the ends of the range of scores.

Running Averages of Frequencies.—As a first illustration of the running-average method, let us apply it to finding the expected frequency f_e in the interval 70–79. The obtained frequency here is 6. We average this along with the two immediately neighboring frequencies, 5 and 14. But we allow the middle frequency to carry twice as much weight; so we add it

twice: $5 + 6 + 6 + 14 = 31$. We have added four numbers; so we divide by 4, obtaining $31/4 = 7.75$. This is our predicted frequency for the interval 70–79. Doing the same for the interval 40–49, we have $7 + 11 + 11 + 4 = 33$. Divided by 4, this becomes 8.25. For the interval 30–39, we have $11 + 4 + 4 + 0$, all divided by 4, which gives us 4.75. If we wish to do so, we may even estimate frequencies in the end classes given, for example, in the interval 20–29. Here we have $4 + 0 + 0 + 0 = 4$, and divided by 4 the outcome is 1.00. All the expected frequencies for this distribution are given in column (3) of Table

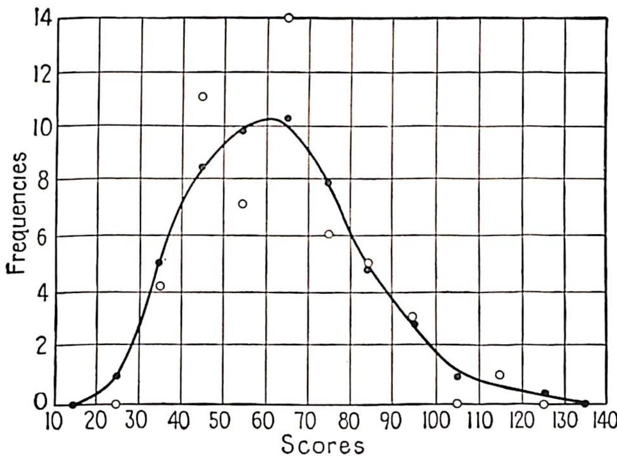


FIG. 3.9.—A smoothed distribution curve for the scholastic-aptitude scores in Table 3.5. The circlets represent obtained frequencies. Dots represent new (smoothed) frequencies obtained by the use of running averages.

3.5. Their sum is equal to 51, which is a rough check upon the accuracy of computation.

Plotting a Smoothed Distribution.—The final step is to plot the smoothed curve, which we have in Fig. 3.9. First the obtained frequencies are plotted as circlets in their proper places. It is always well to show these even though we do not draw the curve through them as before. The expected frequencies are next plotted as points. We can probably see by inspection that the smoothing could be improved upon. In drawing the smoothed curve, we do not feel compelled necessarily to touch all the dots. Being concerned with the general shape freed from probably accidental fluctuations, we take the liberty of further smoothing by inspection and by free-hand drawing. If there were too many irregularities, even in the smoothed points we could, of course, repeat the averaging process, but this is usually not wise, because it tends to flatten the entire distribution too much and should be avoided if possible. In the present instance, very

little further adjustment of frequencies was needed in order to produce the smoothed and rounded contour seen in Fig. 3.9. We may expect with some confidence that the larger population from which this group is drawn will distribute more like the rounded curve than like the irregular one we actually obtained.

A Semigraphic Report in Typewritten Form.[1]—When making reports of frequency distributions, in typewritten form particularly, when the number of cases is not too small, a good form is to let a period stand for one individual, a colon stand for two, and an *x* stand for five, as in Table 3.6.

TABLE 3.6.—A SEMIGRAPHIC REPORT OF TWO FREQUENCY DISTRIBUTIONS

Age at last birthday	Number of men entering college	Number of women entering college
31–35	.	:
26–30	: .	x.
25	x:	x.
24	x.	x:
23	xx.	x:
22	xxxx	x.
21	xxxx:.	xxx.
20	xxxxxxxx	xx: .
19	xxxxxxxxxxxxxxxxx: .	xxxxxxxxx: .
18	xxxxxxxxxxxxxxxxxxxxxxx:	xxxxxxxxxxxxx:
17	xxxxxxxxxxxxx: :	xxxxxxxxxxx:
16	:	x.

When the frequencies are small numbers, the same plan gives an adequate picture if we let an *x* or some other letter stand for each individual.

When Coarse Grouping Is Desirable.—It was indicated in an earlier footnote that there are occasions when the rules given for size and number of class intervals should be modified. In making a graphic representation of data it is often desirable to reduce the number of class intervals, even below 10, and to make the intervals correspondingly larger. Doing so will often provide a much better picture.

In small samples, (*for this particular purpose* we may define a *small sample* as one with an *N* less than 100), with fine grouping, the frequencies are likely to be irregular. Sometimes the effect upon the graphic figure is to produce a "saw-tooth" contour. It is very probable that the population distribution, if we had it, would be smooth and regular. Since we usually want the sample distribution to reflect the general picture of the

[1] I am indebted to H. M. Cox for being introduced to this convenient device, which he attributes to F. S. Beers.

population from which it came and which it is supposed to represent, we would like to avoid those irregularities. One solution already offered is that of smoothing the distribution curve. There are some who object to smoothing as the remedy, and for them there is another possibility. In general, curves will be more regular if grouping is coarser.

Another aspect to this problem is that the particular frequencies we obtain by grouping are strongly dependent upon the choice we make in starting each class interval. With the same size of class interval, we might derive quite a different-appearing frequency polygon simply by making our division points between classes at other places, particularly if the sample is small. One can readily demonstrate this by choosing an appropriate interval of 3, let us say, and by setting up three distributions, starting the lowest interval at 12, 13, and 14, respectively, when the lowest score is 14. By introducing coarser grouping, this phenomenon, too, tends to be counteracted.

Another consideration in this grouping problem is the position of the mode, *i.e.*, the point on the measurement scale corresponding to the highest point on the frequency curve. As different sizes of interval are utilized, and as different starting points for intervals are chosen, so the mode may shift up or down on the measurement scale, even jumping from interval to interval. Coarser grouping will also tend to stabilize the interval and the value of the mode.

Based upon certain mathematical considerations which we cannot go into here, Kelley has proposed that the number of classes to be utilized in the graphic representation of a distribution should be determined roughly from the size of sample as shown in Table 3.7.

From the information given in Table 3.7, one would be justified in using only 8 classes for the ink-blot test data, which have been used so extensively for illustrations in this chapter. This number of classes would mean a class interval of 6, which could, of course, be used, though it is not in the preferred list. An interval of 10, which *is* in the preferred list, would result in only 5 classes, which would be less than are called for in Table 3.7. Remember that the coarser grouping is called for, thus far, only for the purpose of graphic representation. The requirement of 10 or more classes still holds for computations such as we meet in the chapter to follow. Since one is often faced with the need of both graphic and computational use of data, some kind of compromise is practically desirable and defensible in many instances. The illustrative example is probably such an instance. The 10 classes used for the ink-blot data yield a frequency polygon which is rather regular, with one notable inversion, and the same 10-class distribution will serve for the computations required.

TABLE 3.7.—THE NUMBER OF CLASSES TO USE IN PREPARING FREQUENCY DISTRIBUTIONS FOR GRAPHIC REPRESENTATION FOR DIFFERENT SIZES OF SAMPLE*

Sample Size (N)	*Number of Classes*
4– 5	2
6– 8	3
9– 14	4
15– 21	5
22– 32	6
33– 46	7
47– 64	8
65– 89	9
90–117	10
118–153	11
154–192	12
193–255	13
256–315	14

* From Kelley, T. L. *Fundamentals of statistics.* Cambridge: Harvard University Press, 1947. P. 133. Reproduced by permission.

The reader will be reminded in the next chapters, however, that with less than 12 classes, it is necessary to make certain corrections for "grouping errors" when certain accurate computations are desired.

Exercises

1. For each one of the following ranges of measurements, state your judgment of (1) the best size of class interval, (2) the score limits of the lowest class interval, (3) the exact limits of the same interval, and (4) its midpoint.

a. 83 to 197.
b. 4 to 39.
c. 17 to 32.
d. 35 to 96.
e. 0 to 188.
f. −24 to +28.
g. 0.141 to 0.205.

2. Given the following list of scores in a "nervousness" test (Data 3*A*) and using a class interval of 5, set up a frequency distribution. In the first solution, begin the lowest class interval with a score of 35. List all exact limits of class intervals and also exact midpoints. In a second solution, start the lowest class interval with a score of 33. After finishing both solutions, write out a comparison of the two distributions and defend the choice of the one as against the other. As a third solution, use an interval of 3, choosing your own starting places for the classes. Discuss the relative merits of the third distribution as compared with the first two.

DATA 3*A*.—SCORES IN A NERVOUSNESS INVENTORY

59	48	53	47	57	64	62	62	65	57	57	81	83
48	65	76	53	61	60	37	51	51	63	81	60	77
71	57	82	66	54	47	61	76	50	57	58	52	57
40	53	66	71	61	61	55	73	50	70	59	50	59
69	67	66	47	56	60	43	54	47	81	76	69	

3. Given the following list of scores, each of which is the percentage of 400 words judged pleasant by an individual (Data *3B*), set up a frequency distribution making the wisest choice of class interval and class limits.

DATA *3B*.—AFFECTIVITY RATIOS
(All have been rounded to the nearest whole number)

43	62	52	48	46	65	43	48	52	51	57	48	48
38	42	44	46	43	35	42	42	45	44	46	40	40
47	52	38	51	45	38	51	40	46	45	54	55	41
50	59	42	39	56	44	43	47	51	43	50	34	40
53	42	31	44	51	43	48	41	43	48	41	55	

4. Plot a frequency polygon and a histogram for Data *3C*, Group I. State your conclusions about these data as revealed by your plotted distributions.

DATA *3C*.—DISTRIBUTIONS OF CHEMISTRY-APTITUDE SCORES IN TWO FRESHMAN CHEMISTRY COURSES, I AND II

Scores	Frequencies for Group I	Frequencies for Group II
90–94	4	2
85–89	10	0
80–84	14	0
75–79	19	0
70–74	32	2
65–69	31	4
60–64	40	5
55–59	28	12
50–54	29	13
45–49	21	21
40–44	18	21
35–39	10	19
30–34	6	20
25–29	1	14
20–24	3	1
Sums. . . .	266	134

5. Apply the smoothing process described in this chapter to Data *3C*, Group I. Plot a curve based upon the smoothed frequencies but show the original frequencies as points, as was done in Fig. 3.9. In what respects has smoothing changed the picture of these data?

6. Reduce distributions I and II (Data *3C*) to percentage distributions, and plot them on the same diagram. Make a descriptive comparison of the two distributions as drawn.

CHAPTER 4

MEASURES OF CENTRAL TENDENCY

This chapter is about averages, of which there are several kinds. Three of them—the *arithmetic mean* (or *mean*, for short), the *median*, and the *mode*—will be explained here. Two others, the *geometric mean* and the *harmonic mean*, being much less useful to students of psychology and education, will be briefly mentioned.

An *average* is a number indicating the central tendency of a group of observations or of individuals. To the question, "How good is a sixth-grade class in arithmetic?" the most reliable and meaningful kind of answer would be the mean or median in some acceptable test of arithmetical achievement. To the question, "What is the weakest tone to which this dog will respond?" the best kind of answer is to state the average result from a number of trials. In either case a single score or a single measurement of the threshold stimulus would be highly unreliable, for not all measurements, even from repeated observations of the same thing, have the same value. To answer those questions by reciting the long list of individual measurements would be highly uneconomical in the reporting and not very enlightening to the questioner.

The average, whether it be a mean, median, or mode, serves two important purposes. First, it is a shorthand *description* of a mass of quantitative data obtained from a sample. It is surely more meaningful and economical to let one number stand for a group than to try to note and remember all the particular numbers. An average is therefore descriptive of a sample obtained at a particular time in a particular way. Second, it also describes indirectly but with some accuracy the *population* from which the sample was drawn. If the sample of sixth-grade children is representative of all the sixth-grade children in the same school, in the same city, or even in the same county, then the average of their scores tells us much about the average that would be made by the population that they represent, be it school-wide, city-wide, or county-wide. If we examine the dog's hearing under a set of conditions that is characteristic of his general, day-to-day existence, the sample average will be very close to one that we could actually obtain by testing him day after day on many days. It is only because sample averages are close estimates of larger population

averages that we can generalize beyond particular samples at all and make predictions beyond the limits of a sample. This means considerable economy of effort, but far more important than that, it makes possible all scientific investigation. We rarely or never know the average of a population, consequently we do not know by how much our obtained average has missed it, but if our sampling has been done in the proper manner we can estimate about how far we may have missed it, as will be shown in Ch. 9. In the present chapter we will be concerned only with the methods of computing averages from sample data.

The Arithmetic Mean

The Mean of Ungrouped Data.—Most readers already know that to find the arithmetic mean (popularly called the *average*), we sum the measurements and then divide by the number of measurements or cases.

In terms of a formula

$$M = \frac{\Sigma X}{N} \qquad \text{(The arithmetic mean)} \qquad (4.1)$$

where M = arithmetic mean.
Σ = "the sum of."
X = each of the measurements or scores in turn.
N = number of measurements or scores.

In a certain experiment to determine the lowest frequency of vibration of a sound wave that would yield a tone for a human observer, 10 trials were given, with the following results: 13, 17, 15, 11, 13, 11, 17, 13, 11, 11 (cycles per second). The sum of these measurements is 132, and therefore the mean is 13.2 cycles per second. Note that in reporting a mean it is given in terms of the unit of measurement, which is specifically stated. A mean is never an abstract number; it is always a mean of something and is always in terms of some unit of measurement.

As another example, the scores on the ink-blot test found in Table 3.1, when summed, give ΣX equal to 1,480. The mean, with the use of formula (4.1), is

$$M = \frac{\Sigma X}{N} = \frac{1{,}480}{50} = 29.60$$

The mean ink-blot score is 29.60 score units. In practice, it is quite customary in reporting a mean to round to one more figure at the right than the original measurements had—in this case, to keep one decimal

place where the original scores were whole numbers. We report the mean as 29.6 score units.[1]

The Mean of Grouped Data.—When data come to us grouped or when they are too lengthy for comfortable addition without the aid of a calculating machine or when we are going to group them for other purposes anyway, we find it more convenient to apply another formula for the mean:

$$M = \frac{\Sigma fX}{N} \qquad \text{(Arithmetic mean from grouped data)} \qquad (4.2)$$

where the symbols N and Σ have the same meaning as before.

X = midpoint of a class interval.

f = number of cases within an interval.

The solution by way of this formula is illustrated in Table 4.1. Here we

TABLE 4.1.—COMPUTATION OF THE MEAN IN GROUPED DATA

(1) Scores	(2) X Midpoint	(3) f	(4) fX
55–59	57	1	57
50–54	52	1	52
45–49	47	3	141
40–44	42	4	168
35–39	37	6	222
30–34	32	7	224
25–29	27	12	324
20–24	22	6	132
15–19	17	8	136
10–14	12	2	24
Sums.	. .	50 N	1,480 ΣfX

$$\text{Mean} = \frac{\Sigma fX}{N} = \frac{1{,}480}{50} = 29.60$$

have only as many different X values as there are class intervals instead of perhaps as many as there are original measurements. Each class interval has as its X value the midpoint of that interval. This assumes that the midpoint of the interval correctly represents all the scores within that

[1] One could determine the number of accurate significant figures in a mean by applying the rules in Ch. 2 at each step of the operations. Further consideration will be given the question of the number of places to report in a mean after discussion of the standard error of a mean in Ch. 9.

interval. This will not be exactly true in many instances, but the discrepancy is small in any case, and in computing the mean, most of the discrepancies counterbalance others, so that the final result is essentially correct.[1]

In column (2) of Table 4.1, the midpoints of the intervals are given. We must add each midpoint into our total as many times as there are cases within the interval. This means finding for each interval the product f times X, or fX. The fX products are listed in column (4). The sum of the fX products (ΣfX) is equal to 1,480. Dividing this by N, we find the mean to be 29.60, as it was for the same data ungrouped. As was indicated before, we should not be surprised to find a minor discrepancy between the means calculated from grouped and ungrouped data. It happened here that the discrepancy was zero. We may also expect trivial discrepancies in means when the same data are grouped differently, *i.e.*, with different size of class interval or with different starting points for intervals of the same size.

The Mean Computed by the Short Method.—When the original measurements are relatively large numbers, particularly when the midpoints and the frequencies are large numbers, the method just described can well give way to a short-cut procedure that saves pencil-and-paper work. Even greater saving is appreciated when, as in the next chapter, a standard deviation is also to be computed. This procedure requires the use of a guessed average, which we call M', and of new or "coded" values to replace the midpoint values. The steps are illustrated in Table 4.2, including the coding process. In this table it can be seen that many of the actual midpoints would be four-place numbers; for example, the highest interval has a midpoint of 154.5 (midway between 149.5 and 159.5) and consequently the fX products would become also rather large. The coded values for the intervals, given in column (3), are called x' and will now be explained.

Choosing a Guessed Mean.—First we select a guessed mean. This may be chosen anywhere, for its choice is arbitrary. In order to obtain the greatest benefits from the short method, however, it is well to choose a guessed mean rather near to the actual mean, at any rate, somewhere near the center of the distribution. Several criteria guide us in making this choice. One is to place the guessed mean at the midpoint of the middle class interval (if there is an even number of intervals, either of the two middle ones is eligible). The distribution in Table 4.2 is distinctly skewed, however, with the bulk of the cases at the lower part of the range; so the

[1] A discussion of how "grouping errors" affect statistics will be found in the chapter immediately following.

TABLE 4.2.—COMPUTATION OF THE MEAN IN GROUPED DATA BY USING THE SHORT METHOD

(1) Scores	(2) f	(3) x'	(4) fx'	
150–159	2	+6	+12	
140–149	2	+5	+10	
130–139	4	+4	+16	
120–129	1	+3	+ 3	
110–119	5	+2	+10	
100–109	5	+1	+ 5	
				+56
90– 99	12	0	0	
80– 89	10	−1	−10	
70– 79	12	−2	−24	
60– 69	10	−3	−30	
50– 59	1	−4	− 4	
				−68
Sums......	64 N	...	−12 $\Sigma fx'$	

$$c = i\left(\frac{\Sigma fx'}{N}\right) = 10\left(\frac{-12}{64}\right) = -\frac{120}{64} = 1.88$$

$$M = M' + c = 94.5 + (-1.88) = 94.5 - 1.88 = 92.62$$

mean we find will probably fall in an interval lower than the middle one. Another criterion is to choose the interval containing the median (see Table 4.3 for the method of finding the median). In this distribution, the median falls within the interval for scores 80–89. This is farther from the center than we would ordinarily go for the guessed mean. Another guide is to choose an interval that has a large number of cases—in fact, the largest number. Here such an interval is that for scores 90–99. As a good compromise among all of these criteria, the interval labeled 90–99 seems best. We should actually come out with the same computed mean no matter which interval we chose for the guessed mean; the choice is dictated entirely by the desire to keep the numbers small so that "headwork" can replace paper-and-pencil work as much as possible.

The Size of Class Interval Becomes the Temporary Working Unit.—Having chosen the interval 90–99, we guess the mean to be at the midpoint of this group, the midpoint being 94.5 (midway between 89.5 and 99.5). The score point of 94.5 becomes the temporary zero point for our measuring scale. In column (3), a zero is written in line with the interval whose midpoint is 94.5. The first interval above is given a value of +1; the second,

+2; the third, +3; etc. The first interval below is given a value of −1; the second, −2; etc. These x' values now represent the class intervals, which are just one unit apart. The new unit is equivalent to 10 score units, a fact that we shall have to remember later.

The Correction to Add to the Guessed Mean.—From here on, the steps are similar to those taken in Table 4.1. Next we find the fx' product for each interval, *taking great care to record algebraic signs.* All products above the guessed mean are positive, and all products below are negative. The sum of the positive products is +56, and the sum of the negative products is −68. The algebraic sum of the entire column is therefore 56 − 68, which equals −12. The $\Sigma fx'$ therefore equals −12. From this we can find directly how far the actual mean is from our guessed mean. The actual mean is equal to M' plus a correction c, and this correction is given by the formula

$$c = i\left(\frac{\Sigma fx'}{N}\right) \qquad \text{(Correction to add to the guessed mean)} \qquad (4.3)$$

where i = size of the class interval.
x' = deviation of a class interval from the guessed mean in terms of i as the unit.
f = frequency within a class interval.
N = total number of measurements.

In this problem, $i = 10$, $\Sigma fx' = -12$, and $N = 64$. Therefore

$$c = 10\left(\frac{-12}{64}\right) = -\frac{120}{64} = -1.88$$

Adding this correction to the guessed mean, we have

$$M' + c = 94.5 - 1.88 = 92.62$$

The mean is 92.62 score units, but we should report it merely as 92.6 score units.

A Summary of the Short Solution of the Mean.—The steps involved in the short method of computing the mean may be summarized as follows:

Step 1. Set up the frequency distribution.

Step 2. Choose a guessed mean. This is the midpoint of the interval (1) near the center of the distribution; or (2) containing the median or mode or both; or (3) probably containing the actual mean.

Step 3. Assign to the class intervals new small integral values, starting with zero at the interval containing the guessed mean, with posi-

tive values above and negative values below. Call these new values x'.

Step 4. Find the fx' product for each interval, and record in a column.
Step 5. Sum the fx' products algebraically. This is $\Sigma fx'$.
Step 6. Divide the sum of the fx' products by N.
Step 7. Multiply this quotient by i, the size of the class interval. This gives the correction c.
Step 8. Add this correction algebraically to the guessed mean. This gives the mean.

A single formula representing the preceding steps is

$$M = M' + i\left(\frac{\Sigma fx'}{N}\right) \quad \text{(Arithmetic mean from grouped and coded data)} \qquad (4.4)$$

where the symbols are as previously defined.[1]

The Median

The *median* is defined as that point on the scale of measurement above which are exactly half the cases and below which are the other half. Note that it is defined as a *point* and not as a score or any particular measurement. If this conception is kept clearly in mind, many difficulties will be forestalled. Some textbooks on statistics give a different definition of median for ungrouped as compared with grouped data and recommend two different procedures for computing the median. Here we shall apply the same definition to both cases and be consistent in computation throughout.

The Median from Grouped Data.—It is probably easier to grasp the process of computing a median in grouped data. For a first illustration, consider Table 4.3. Here there are 28 cases; so the median is that number of points on the measuring scale above which there are 14 cases and below which there are 14. Counting frequencies from the bottom upward, we find that $4 + 1 + 1 + 10 = 16$ cases, or 2 more than we want. To make 14 cases, we need 8 out of the 10. The median lies somewhere within the interval 15–19, whose *exact* limits are 14.5 and 19.5. We assume for the sake of computation that the 10 cases within this interval are evenly spread over the distance from 14.5 to 19.5 (see Fig. 4.1). We must interpolate within this range to find how far above 14.5 we need to go in order to include the 8 cases we need below the median. We must go 8/10 of the way, for 8 is the number we require, and 10 is the total number in the

[1] The solution by use of formula (4.4) will be better understood by those who follow the proofs offered in Appendix A and who apply those proofs here.

TABLE 4.3.—COMPUTATION OF THE MEDIAN SIZE OF CLASS IN A CERTAIN SCHOOL, WITH THE USE OF GROUPED DATA

Class size	f	
40–44	1	
35–39	0	
30–34	3	
25–29	5	
20–24	3	12 = number of cases above the interval containing the median
15–19	10	
10–14	1	6 = number of cases below the interval containing the median
5– 9	1	
0– 4	4	
$N = 28$		

$$Mdn = 14.5 + 8/10 \times 5 = 14.5 + 4.0 = 18.5$$
$$Mdn = 19.5 - 2/10 \times 5 = 19.5 - 1.0 = 18.5$$

interval. The total distance is 5 units; so on the scale of measurement we go 8/10 of 5, or exactly 4.0 units. Adding this 4.0 to the lower limit of the class interval 14.5, we get $14.5 + 4.0 = 18.5$ as the median.

We can check this by counting down from the top of the distribution until we include $N/2$ of the cases; 14 in this problem. Starting at the top, we find that

$$1 + 0 + 3 + 5 + 3 = 12.$$

We need 2 more cases out of the next group of 10. We must go 2/10 of the way below the *upper* limit of the interval, that is, below 19.5. This means 2/10 of 5 or exactly 1.0 unit. The upper limit, 19.5 minus 1.0, gives us 18.5 for the median, which checks with the one obtained by counting up from below. It is well always to check the determination of a median in this manner, and to do so involves very little work. If the two estimates do not agree exactly, something is wrong.

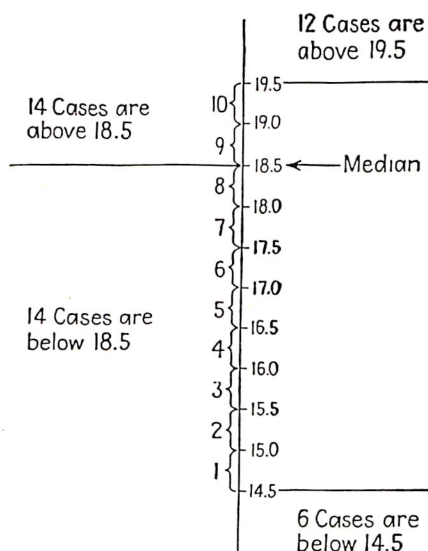


FIG. 4.1.—Showing how the 10 cases in the interval 14.5 to 19.5 are distributed. Each case is assumed to occupy a tenth of the interval, or one-half of a score unit. The eighth one extends up to the point 18.5, which is the median.

TABLE 4.4.—COMPUTATION OF THE MEDIAN SCORE IN A SENTENCE-CONSTRUCTION TEST AS GIVEN TO 37 MEN

Scores	f	
37–38	1	
35–36	2	
33–34	0	
31–32	1	
29–30	0	
27–28	6	
25–26	5	15 = number of cases above interval containing the median
23–24	8	
		14 = number of cases below interval containing the median
21–22	8	
19–20	5	
17–18	1	
$N = 37$		$N/2 = 18.5$

$$Mdn = 22.5 + \frac{4.5}{8} \times 2 = 22.5 + \frac{9}{8} = 22.5 + 1.125 = 23.6$$

$$Mdn = 24.5 - \frac{3.5}{8} \times 2 = 24.5 - \frac{7}{8} = 24.5 - .875 = 23.6$$

To take another example with grouped data, consider Table 4.4, where N is an odd number. Here $N/2$ is 18.5, but the principle of interpolating within an interval for the exact median is just the same. Counting up from below, we find that $1 + 5 + 8 = 14$, which lacks 4.5 cases of including the lower half. In the next interval, we must go 4.5/8 of the way, or 4.5/8 times 2, which equals 9/8, or 1.125. Adding this many units to the lower limit of the interval (22.5), we have 23.625 as the median; or dropping all but one decimal place, we report the median as 23.6 score units. Checking by counting down from the top, we find 15 cases above the point 24.5. Going 3.5/8 of the way down into the interval of 2 units, we find that we must deduct 0.875 from 24.5 to find the median. When rounded to one decimal place, the median is 23.6, as before. In terms of a formula, the interpolated median is found from below by

$$Mdn = l + \left(\frac{\frac{N}{2} - F_b}{f_p} \right) i \quad \text{(Interpolation of a median from below)} \qquad (4.5a)$$

where l = exact lower limit of class interval containing median.
F_b = sum of all frequencies below l.
f_p = frequency of the interval containing Mdn.
N and i are defined as usual.

In terms of a similar formula, the median is found from above by

$$Mdn = u - \left(\frac{\frac{N}{2} - F_a}{f_p}\right)i \quad \text{(Interpolation of a median from above)} \qquad (4.5b)$$

where u = exact upper limit of the interval containing the median.
F_a = sum of all frequencies above u.
Other symbols are as defined previously.

A Summary of the Steps for Interpolating a Median.—The steps for computing a median from grouped data may be summarized as follows:

Step 1. Find $N/2$, or half the number of cases in the distribution.
Step 2. Count up from below until the interval containing the median is located.
Step 3. Determine how many cases are needed out of this interval to make $N/2$ cases.
Step 4. Divide this number needed by the number of cases within the interval.
Step 5. Multiply this by the size of class interval.
Step 6. Add this to the exact lower limit of the interval containing the median.
Step 7. Check by adding down from the top to find to what point the upper half of the cases extend in a manner analogous to that described in Steps 2 to 5 inclusive.
Step 8. Deduct the number of score units found in Step 7 from the exact upper limit of the interval containing the median.

Some Special Situations.—There are some instances in which things do not turn out just as they did in the two illustrative examples.

When the Median Falls between Intervals.—If it should happen, in adding up cases from below, that half the cases take in *all* the cases in the last interval, the median is then the exact upper limit of that interval. In counting down from above, it would be found that all the cases in the interval just above this one would also be required to make $N/2$; so its exact bottom limit would be the median. This coincides with the exact upper limit of the interval below; thus, the median checks. As an example, note the following fictitious data:

Scores	20–24	25–29	30–34	35–39	40–44	45–49	50–54	55–59
f	2	7	10	15	18	8	3	5

Here $N/2$ is 34. This many cases takes us exactly through the interval 35–39. The median is 39.5. From above down, we are carried through the interval 40–44, whose lower limit is 39.5. Again the median is 39.5.

When There Are No Cases within the Interval Containing the Median.—Another question arises when the median falls within an interval where there are *no* cases. It is even possible that in the region of the median, two or more intervals have frequencies of zero. If the range having no cases is one interval, the median may be taken as the midpoint of that interval, but this gives a very crude estimate unless the size of the interval is small—for example, not over three units. If that range covers two or more intervals, no good estimate can be made for the median.

Scores	5–7	8–10	11–13	14–16	17–19	20–22	23–25	26–28
f	1	7	9	0	6	7	2	2

In the data just preceding, the median is 15.0, which is midway between 13.5 (to which point the lower half of the cases extend) and 16.5 (to which point the upper half of the cases extend). Or it is the arithmetic mean of those two limits, for 16.5 + 13.5 divided by 2 is 15.0.

The Median from Ungrouped Data.—Things learned in finding a median in grouped distributions should carry over almost intact to the use of ungrouped data. The median is a *point* on the measuring scale. In ungrouped data, each score or measurement is assumed to occupy a *range* of one unit. The median either falls within one of those units or somewhere between units. The first step is to arrange the measurements in order of their size. The list of 10 measurements of the threshold for pitch as given on page 59, when placed in rank order, becomes

11, 11, 11, 11, 13, 13, 13, 15, 17, 17

As in the case of grouped data, it is assumed that the four 11's occupy the range from 10.5 to 11.5; the three 13's occupy the range from 12.5 to 13.5, etc. Counting from below to include 5 cases brings us to the first 13 that must be included among the 5. We must therefore extend 1/3 of the way in the interval of 1 unit, or 0.33 unit into the interval, starting at 12.5. The median is 12.5 + 0.33, which equals 12.83, or, when rounded, 12.8. In checking from above, the median is found at 13.5 − 0.7, which also equals 12.8.

In the series of measurements

2, 5, 7, 8, 9, 10, 17

the median comes midway in the fourth one, which is 8. Since 8 occupies a range of 7.5 to 8.5, the median is the midpoint of this range, or exactly 8.0. In the series of measurements

7, 9, 10, 12, 13, 15, 18, 20

four are 13 or above, and four are 12 or below. The division between upper and lower halves comes at 12.5, which is the median in this case. In the array of scores

15, 17, 18, 20, 23, 24, 27, 30

the lower half extends up to 20.5, and the upper half extends down to 22.5. Midway between these two values is the point 21.5, or the average of the two.

It is probably obvious that the median of so small a number of observations cannot be very reliable, and we should not place too much reliance upon it or carry our calculations to more than one decimal place (we might even report nearest whole numbers); but in order to keep consistent certain principles of the median and of the process of computing it, certain steps have been emphasized. Whenever there is doubt concerning special cases not covered in these illustrations, an application of these principles should take care of the matter.

The Mode

The *mode* is strictly defined as the *point on the scale of measurement with maximum frequency in a distribution.* When we have ungrouped data, the mode is that measurement which occurs most frequently. Usually it is somewhere near the center of the distribution, and in a strictly normal (Gaussian) distribution it coincides with the mean and the median.

The Crude Mode.—*In a distribution of grouped data, the crude mode is the midpoint of that class interval having the greatest frequency.* In Table 4.1, the highest frequency is 12, for the interval 25–29. The midpoint of this interval is 27; so the mode is taken to be 27.0. In Table 4.2, there are two intervals with the same maximum frequency of 12. If these two intervals had been separated by more than one intervening interval of lower frequency, we should be justified in saying that the distribution is *bimodal* (having two modes). But the single intervening frequency of 10 hardly gives us sufficient basis for this conclusion. The distribution is therefore probably really unimodal, but we are not able to decide upon its crude mode. A calculated mode can be found, as we shall soon see.

In Table 4.3, the crude mode is clearly 17.0. In Table 4.4, the maximum frequency is shared by two neighboring intervals. In a situation

like this, we do the reasonable thing of assigning the crude mode to the dividing point between these intervals, which is 22.5. Unless the data are reasonably numerous, so that there is clearly an interval of highest frequency, we should not attempt to assign a modal value to the distribution. For example, the 10 measurements of threshold for pitch present an unusual situation with the greatest frequency (four cases) of 11, which is at one end of the distribution. Following right behind is the measurement 13, with three cases. Here it would be rather meaningless to say that the mode is 11.

Estimation of the Mode by Coarse Grouping.—In certain methods of estimating the mode (such as the interpolation method described below) it is frequently helpful to resort to coarser grouping (smaller number of class intervals) than usual. This results in larger frequencies within the classes. Larger frequencies are more stable in the sense that a change of one or two cases more or less would affect them relatively less than when frequencies are small. They would change relatively less, also, from sample to sample drawn from the same population in the same manner. Furthermore, differences between frequencies in different classes are likely to be greater so that there is less doubt as to which interval contains the mode. Following a recommendation made by Kelley,[1] the optimal conditions for estimating the mode prevail when the numbers of classes are as given in Table 4.5.

TABLE 4.5.—OPTIMAL NUMBERS OF CLASSES FOR ESTIMATING THE MODE FOR DIFFERENT SIZES OF SAMPLE

N	20	30	60	125	200	400	1300
Classes......	3	4	5	6	8	10	13

Interpolation of the Mode.—In the method given above for the estimation of the crude mode, the midpoint of the interval with the largest frequency was chosen as the best value. This is adequate for smaller samples (less than 100 cases) and even for larger samples when the distribution is symmetrical (not skewed). When distributions are noticeably skewed, however, the midpoint value is not as accurate an estimate as we can make. With larger samples, and even with smaller ones that yield a regular contour, we can interpolate within the modal interval to obtain a more accurate estimate of the mode. The procedure is best explained by reference to Fig. 4.2.

[1] Kelley, T. L., Fundamentals of statistics. Cambridge: Harvard University Press, 1947. P. 259. In Table 4.5, the relation of the range of measurements to the standard deviation, particularly in small samples, has been taken into account (see Table 5.8).

In that illustration only the intervals with the three highest frequencies are shown. The three highest frequencies should be neighboring if this procedure is used. Nor should this procedure be used if these three frequencies do not reflect a skewing in the same direction as that for the entire distribution. The three in this illustration imply negative skewing for the total distribution.

The principle of this method is to fix the value of the mode within the interval in proportion to the frequencies in intervals on either side of the

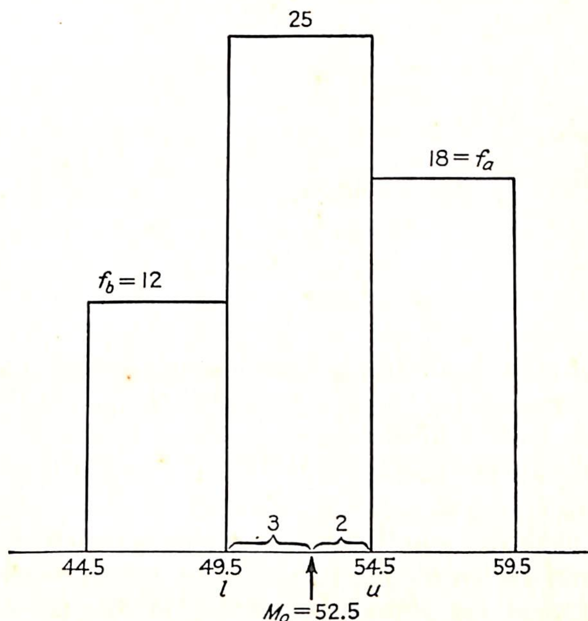


FIG. 4.2.—Illustration of the estimation of a mode from the three frequencies nearest to it.

modal interval. Here the greater neighboring frequency is in the interval above the modal one, consequently the mode is estimated toward the upper limit of the modal one. In the interpolation process we assume that the nearness of the mode to the lower or upper limit of the modal interval should be proportional to the frequency in the interval neighboring upon each limit. The modal point can be estimated by the formula

$$Mo = l + \left(\frac{f_a}{f_a + f_b}\right) i \qquad \text{(Interpolated mode)} \qquad (4.6a)$$

where l = exact lower limit of modal interval.
f_a = frequency in interval immediately above.
f_b = frequency in interval immediately below.
i = size of class interval.

The computation can be checked by working from the upper limit, by the formula

$$Mo = u - \left(\frac{f_b}{f_a + f_b}\right) i \qquad \text{(Another interpolation of the mode)} \qquad (4.6b)$$

Applying these formulas to the data represented in Fig. 4.2, we have, first,

$$\begin{aligned} Mo &= 49.5 + \left(\frac{18}{18 + 12}\right) 5 \\ &= 49.5 + .6 \times 5 \\ &= 49.5 + 3.0 \\ &= 52.5 \end{aligned}$$

Second, we have

$$\begin{aligned} Mo &= 54.5 - \left(\frac{12}{18 + 12}\right) 5 \\ &= 54.5 - .4 \times 5 \\ &= 54.5 - 2.0 \\ &= 52.5 \end{aligned}$$

It will be seen that the result is conservative and that it would require an unusually uneven balance between f_a and f_b to move the mode very far from the midpoint of the modal interval. This is probably as it should be, but it suggests that with f_a and f_b very similar there is little need for applying the method.

If the contour of the distribution is irregular, a preliminary step to the use of this interpolation method would be to introduce some smoothing by the running-average procedure described in the preceding chapter. Enough regularity would then be introduced to make feasible the use of formulas (4.6*a*) and (4.6*b*).

The Mode Estimated from the Mean and Median.—Fortunately, because of certain mathematical relationships between the mode and the other two measures of central tendency, we can estimate the mode from them. A simple approximation formula is

$$Mo = 3Mdn - 2M \qquad \text{(Estimation of a mode from mean and median)} \qquad (4.7)$$

In other words, the mode equals three times the median minus two times the mean.

Applying this formula, we can now estimate the mode of the distribution in Table 4.2, in which we were unable to decide upon a crude mode. The median for this distribution is 88.5, and the mean is 92.62. Although we rounded the mean to one decimal place in reporting it, in further cal-

culations with it, we do well to keep the second decimal place. Applying formula (4.7), the computed mode equals

$$(3 \times 88.5) - (2 \times 92.62) = 265.5 - 185.24 = 80.26.$$

Rounded to one decimal place, the estimated mode is 80.3. Reference to the distribution in Table 4.2 again will show that this point comes about midway among the four high frequencies. Had we done a very reasonable thing and placed the crude mode midway among these four intervals, it would have been at 79.5, which is less than one unit from the calculated mode.

The mean of the distribution in Table 4.3 is 19.14 and the median is 18.5. The calculated mode is $(3 \times 18.5) - (2 \times 19.14)$, which equals $55.5 - 38.28$, or 17.22. This is separated from the crude mode, which is 17.0, by a trivial amount. In the distribution in Table 4.4, the median is 23.6, and the mean is 24.52. From this information, the mode is estimated as 21.8, which deviates from the crude mode only 0.7 unit. It may add meaning to the computed mode to say that it is the point on the measuring scale at which the smoothed distribution curve probably has its highest point.

When to Employ the Mean, Median, and Mode

Certain Advantages of the Mean.—The arithmetic mean is to be preferred whenever possible because of several desirable properties. In the first place, it is generally the most reliable or accurate of the three measures of central tendency. By this we mean that from sample to sample of the same population, the mean will ordinarily fluctuate less widely. Another reason is that the mean is better suited to further arithmetical computations. Deviations of single cases from the central tendency are important information about any distribution. Much is done with these deviations, as will be seen in the following chapter. It will also be found that we square those deviations, and this we are really justified in doing only when the deviations are taken from the mean. When distributions are reasonably symmetrical, we may almost always use the mean and should prefer it to the median and mode. On the other hand, there are instances, particularly when distributions are skewed and when the mean would lead to erroneous ideas about a distribution, in which other measures of central tendency are better used.

A Comparison of the Mean with Median and Mode.—One property of the mean is that it is sensitive to the size of extreme measurements when they are not balanced by other extreme measurements on the other side

of the middle. In the following set of measurements, the mean is 9 and the median is 9:

4, 5, 7, 9, 11, 13, 14

Now, if the 14 had been 23 instead of 14, the median would be unchanged, but the mean would become 10. There are still an equal number of cases above and below 9. So far as the median is concerned, the 11, 13, and 14 could have been 110, 130, and 140, and still the median would be 9. But in this rather unusual but not impossible event, the mean would become 57.9, where formerly it was only 9. The conclusion to be drawn is that when, in a small sample particularly, there are any very extreme measurements not balanced by other extreme measurements in the other direction, the median is to be preferred to the mean.

Some Mathematical Properties of the Arithmetic Mean and the Median.—A better appreciation of the nature of the mean and of the median may be gained by noting some of their mathematical peculiarities. To illustrate, let us use the data presented in Table 4.6. There six scores are given for six individuals. The mean of these scores is 6.0 and the median is 4.5.

TABLE 4.6.—ILLUSTRATION OF CERTAIN PROPERTIES OF THE ARITHMETIC MEAN AND THE MEDIAN

(1) Person	(2) Score	(3) Deviations from the mean	(4) Deviations from the median	(5) Deviations from the mean, squared	(6) Deviations from the median, squared
A	2	−4	−2.5	16	6.25
B	3	−3	−1.5	9	2.25
C	4	−2	−0.5	4	0.25
D	5	−1	+0.5	1	0.25
E	9	+3	+4.5	9	20.25
F	13	+7	+8.5	49	72.25
Sums..........	36	0	+9.0	88	101.50
Means.........	6.0	0.0	+1.5		
Median........	4.5	—	—		

The first feature to be pointed out is that the mean is *the center of gravity* of the scores. In Fig. 4.3 we have the six scores represented on the measurement scale. Imagine that the six individuals are arranged in their proper places along this scale. Imagine that the scale itself is a rigid plank or bar. The six persons may be regarded as exactly the same in all respects except for their scores on this scale. Each "weighs" the same; his effect upon the tilting of the bar depends only upon his position upon it.

If we wish to rest the bar upon a single fulcrum in such a position that the bar will be perfectly balanced, that position must coincide with the mean. The measurements in any sample are perfectly balanced about the arithmetic mean.

Each individual in this small distribution carries an effective weight in proportion to his distance from the mean. In the parlance of the physicist, each person's distance from the mean is called a *moment*. In statistics, also, we often speak of moments in a similar sense. In column (3) of Table 4.6, each of the six moments for this small distribution is given. They are more commonly called *deviations from the mean* or simply *deviations*. The size of each deviation indicates how much effective weight the moment carries and its algebraic sign tells in what direction that weight is applied. The algebraic sum of these moments is zero, as it always is when the arithmetic mean and the deviations are correctly computed. This is

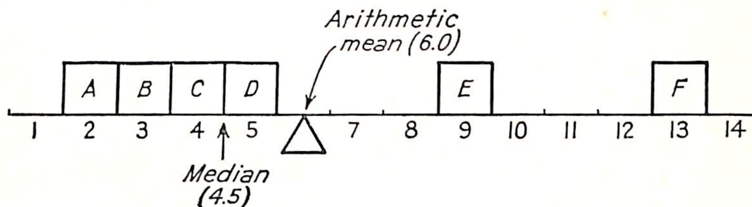


FIG. 4.3.—Illustration of the positions of six cases with respect to the arithmetic mean and with respect to the median. If all cases carry equal weight, they are perfectly balanced when the fulcrum is placed at the arithmetic mean.

simply another indication that the mean is a center of gravity, for the positive and negative moments about the mean are perfectly balanced.

The arithmetic mean is the only value in a distribution from which the deviations always sum algebraically to zero. To show that the median does not qualify in this respect, let us find the deviations of the six scores from the median and sum them (see Table 4.6). The algebraic sum of the deviations from the median is 9.0. This means a net balance of 9 units on the plus side. A fulcrum placed at the point 4.5 on the scale would be seriously overbalanced toward the end with the high scores. This comes from the fact that in computing a median we ignore the distance of each case from the central value. If we want the bar to balance when the fulcrum is placed at the median value, we will have to rearrange the cases, treating all cases above the median as if they had the same value and all cases below the median as if they also had the same value and a value as far below the median as the above-median group was placed above it.

Not only are the deviations from the mean balanced about it but they have another important property. If we square each deviation, we have

the squared moments about the mean. The peculiarity of the mean is that the sum of the squared deviations about it is smaller than that for the squared deviations about any other value. In most of the following chapters we will be concerned with squared deviations from the mean. For the present, it is merely significant to point out that when squared deviations are considered the arithmetic mean is closest to the measurements of the sample as a whole. In Table 4.6 we can see that for this small sample the sum of squared deviations is much smaller when the reference point is the mean than when it is the median, the two sums being 88 and 101.5. The reader may verify the fact that 88 is the smallest possible sum of squared deviations in this sample by arbitrarily choosing other values as possible points of central tendency.

Central Tendencies in Skewed Distributions.—In skewed distributions, the mean is always pulled toward the skewed (pointed) end of the curve,

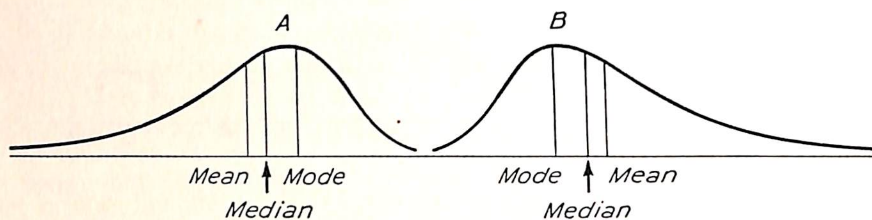


Fig. 4.4.—Two skewed distributions, (*A*) skewed negatively and (*B*) skewed positively, showing the relative positions of modes, medians, and means. Note that the mean is displaced farther from the mode toward the skewed end of the distribution and that the median is displaced about two-thirds as far.

as Fig. 4.4 shows. The arithmetic mean, as the center of gravity of the distribution, is weighed toward the extreme values, as was demonstrated above. The *sum* of the deviations on the one side of it equals the *sum* of the deviations on the other side. The median comes at a point that divides the area under the distribution curve into two equal parts. The *number* of scores on the one side of it equals the *number* of scores on the other. The interpretations of mean and median should be made accordingly. For example, for the data on class size in Table 4.3, the *median* of 18.5 tells us that half of the classes had 19 or more students enrolled and half of them had 18 or less. The mean class size, which is 19.1, tells us that if all the enrolled students had been reapportioned so as to make all classes the same size, the enrollment in each class would have been 19.1, or 19, with a few students left over.

When the Mean Is Misleading.—In some instances, to give the mean of a distribution only is highly misleading; for example, in a study of class size in a certain university, among 62 classes, there were 2 classes having

more than 200 students, and 2 having between 100 and 200 students, all the remaining classes except 2 being smaller than 60. The average size of the 62 classes was 34, but this was not very typical, because half of the classes had 20 or less (the median was 20.5). The most *typical* size of class would be given as the *mode*, which was 17 (crude mode). If our purpose happened to be to equalize the size of classes, assuming that this were practical, we could conclude that there would be 34 students per class. If we wanted to decide as a matter of educational policy whether or not there were too many small classes in general and if we had concluded beforehand that most teachers can successfully handle 30 students in a group, then the median would tell us, without knowing anything more about the distribution, that there were entirely too many small classes. The mean would not have told us this, because it was higher than 30. If we were piloting a visiting inspector about the buildings while classes were in session and wished to prepare him for the most likely size of class he would find at random, we should give him the mode, since this size is more likely to occur than any other one size. If we were purchasing equipment to suit classes of various sizes, we should adapt it, if necessary, most often to classes of modal size, though in this case we should also want to know more about the entire frequency distribution.

Mean and Median Often Both Reported.—In reporting upon central tendencies of skewed distributions, it is usually well to state both the mean and the median, since each tells its own story, and from the difference between the two we can immediately infer in what direction the distribution is skewed and about how strongly. Although the mode is easily and quickly determined and will often serve until better averages can be computed, it should probably never be reported alone and need not be reported with the other two averages except when it is meaningful to do so. When a distribution is symmetrical about the mode, the three averages will coincide, and so only one of them, preferably the mean, need be reported, together with the fact that the distribution is symmetrical.

When the Median Is Especially Called For.—There are one or two kinds of distribution in which the median is the only satisfactory average.

Distributions with Indeterminate Values.—There are some distributions in which some of the extreme values are not accurately determined. We know that they lie out beyond a certain point on the scale but we do not know just how far. In certain work-limit tests, for example, some subjects would work on for unusual lengths of time if permitted to do so. Suppose that all those who work on a certain test up to 10 min. are arbitrarily stopped. They are in the minority, so a median can be found. Time spans up to 10 min. may be classified as usual into chosen class

intervals. From 10 min. up, we find the laggards grouped together. We do not know just how long they might have kept working had we let them continue. An arithmetic mean cannot be determined here, but median and mode can still be utilized.

When Equality of Unit Is Uncertain.—In another instance, we are not sure that all the units of our measuring scale are equal. This is particularly true in the psychological scaling methods of rank order and of equal-appearing intervals. In the former case, a number of judges have placed several objects or persons in rank order for some quality. Though the ranks are numerically equidistant, the things ranked probably are not. When combining ranks for any one object, we do less violence to the measurement if we find a median rather than a mean. In the other instance, though objects are placed in piles or categories that seem equidistant to the observer, again we are not sure that his categories are numerically equidistant, and the median is a safer statistic to compute. It is also true in this scaling method that distributions of judgments for objects very high or very low on the scale are skewed or even truncated because of the "end effect." By the end effect, we mean that although some judges would like to place some stimuli above the highest pile or category (or below the lowest), they are not permitted to do so. Some objects or persons rated thus pile up in the end categories when some of these times they should have gone beyond the end. This fact will distort the arithmetic mean but will not influence the median so long as not more than half of all the judgments for an object fall in the end group.

A Summary of When to Use the Three Averages.—In brief, the following rules will generally apply:

1. *Compute the arithmetic mean when*
   a. The greatest reliability is wanted. It usually varies less from sample to sample drawn from the same population.
   b. Other computations, as finding measures of variability, are to follow.
   c. The distribution is symmetrical about the center, particularly when it is approximately normal.
   d. We wish to know the "center of gravity" of a sample.
2. *Compute the median when*
   a. There is not sufficient time to compute a mean.
   b. Distributions are badly skewed. This includes the case in which one or more extreme measurements are at one side of the distribution.
   c. We are interested in whether cases fall within the upper or lower halves of the distribution and not particularly in how far from the central point.

d. An incomplete distribution is given.
e. There is uncertainty about the equality of the unit of measurement.

3. *Compute the mode when*
   a. The quickest estimate of central tendency is wanted.
   b. A rough estimate of central tendency will do.
   c. We wish to know what is the most typical case.

MEANS IN SOME SPECIAL SITUATIONS

The measures of central tendency described thus far will take care of the great majority of situations in which such statistics must be computed. There are some problems, which, though rare, require other treatment. Four of these will be briefly mentioned: means of arithmetic means, means of percentages (and proportions), geometric means, and harmonic means.

Finding Means of Arithmetic Means.—When one has the means of several samples, presumably from the same population, on the same test or scale, he may want to know the over-all mean for the samples combined. At first thought, it might seem appropriate simply to average the several means just as one would average single observations. This would be proper procedure provided the samples are of the same size. If the *N*'s in the samples differ, however, the means are not equally reliable. In order to extract the best information about the central tendency of the entire sample, we should weight each mean according to the number of cases in the sample from which it was derived, for a mean's reliability is in proportion to the size of sample. This procedure is equivalent to pooling all the single measurements from the different samples and computing a single over-all mean. We can accomplish the same end by computing a weighted mean of the means which we already know. The general formula for computing a weighted mean is

$$ {}_wM = \frac{\Sigma WX}{\Sigma W} \qquad \text{(A weighted arithmetic mean)} \qquad (4.8)$$

where ${}_wM$ = weighted mean.
W = weight.
ΣWX = sum of the values being averaged, each multiplied by its appropriate weight.
ΣW = the sum of the weights.

Table 4.7 illustrates the application of this formula. In the problem represented there, four means differing considerably had been derived from samples ranging from approximately 400 to approximately 2,700

TABLE 4.7.—COMPUTATION OF A MEAN OF ARITHMETIC MEANS, WITH AND WITHOUT WEIGHTING THE SAMPLES*

(1) Group	(2) Number in the sample $N_i = W$	(3) Mean of the sample $M_i = X$	(4) Weighted mean $N_iM_i = WX$
A	15	25.6	384.0
B	27	31.3	845.1
C	9	38.7	348.3
D	4	32.5	130.0
Sums......	$55 = \Sigma W$	$128.1 = \Sigma X$	$1707.4 = \Sigma WX$
Means.....	....	$32.0 = M_x$	$31.0 = {}_wM_x$

* The samples were of scores on a perceptual-speed test administered to aviation students and other military personnel. The sizes of samples were approximately 100 times the values given. Rounding was done to simplify the illustration. It probably did not affect the size of the weighted mean materially. N_i is the number of cases in sample *I*, and M_i the mean of sample *I*.

cases each.[1] The unweighted mean of these four means would be 32.0, whereas the weighted mean is 31.0. The latter is much more representative of *all* the individuals in the combined sample.

When the means to be averaged are very close together, as they will ordinarily be when samples are drawn from the same population and are not too small, and when the N's do not vary much from sample to sample, the weighted and unweighted means will be very close together. In certain situations, then, the unweighted mean may be reported. But if the composite mean is to be used for further computations, in which case it should often be estimated to the second decimal place, weighting certainly is called for.

The Mean of Percentages or of Proportions.—The weighting procedure just described is even more important in determining the mean of a series of percentages or of proportions. Table 4.8 illustrates this point. The data in that table have to do with the percentage of pilot students eliminated in certain schools during one training period. Had the schools had the same enrollment, or even very nearly the same, the unweighted mean would suffice. Since the largest class is nearly four times as great as the smallest, however, and since elimination rates vary from 3.3 to 27.2, there is a marked difference between weighted and unweighted means. If we wished to know the over-all elimination rate in order to make decisions for some administrative purpose, the unweighted mean would

[1] The means are so different and samples are so large that it is highly unlikely that the samples came from the same populations. They will serve to illustrate the procedure nevertheless.

TABLE 4.8.—COMPUTATION OF AN AVERAGE PERCENTAGE*

(1) School	(2) Number enrolled N_i	(3) Number eliminated $N_iP_i/100$	(4) Per cent eliminated P_i
G	243	55	22.6
H	63	7	11.1
K	196	43	21.9
L	61	2	3.3
S	125	34	27.2
Sums......	688 = ΣN_i	141 = $\Sigma N_iP_i/100$	86.1 = ΣP_i
Means.....	137.6 = M_N		17.2 = M_p†

* The data represent students enrolled in five AAF pilot schools selected to illustrate this procedure.

† The weighted mean of the percentages equals 14,100/688 = 20.5. The value 17.2 is the unweighted mean.

be misleading. Certainly, when the percentage or the proportion in a composite is wanted for further computations, the weighting procedure is essential, unless the sample N's are exactly equal.

In terms of a formula, the weighted mean of a percentage is

$$_wM_p = \frac{\Sigma N_iP_i}{\Sigma N_i} \qquad \text{(Mean of percentages where } N\text{'s differ)} \qquad (4.9)$$

where N_i = number in each sample.
P_i = percentage for each sample.
ΣN_iP_i = sum of products of each percentage times its corresponding N.
ΣN_i = sum of the sample N's.

A completely analogous formula applies to finding the weighted mean of proportions, in which case p is substituted for P.

The Geometric Mean.—The *arithmetic* mean of two numbers is found by adding them and dividing by two. The *geometric* mean of two numbers is found by *multiplying* the two numbers and then taking the *square root.* The arithmetic mean of 2 and 18 is 10.0. The geometric mean is

$$\sqrt{2 \times 18} = \sqrt{36} = 6.0.$$

The geometric mean of three numbers is the cube root of their product; of four numbers the fourth root of their product; and so on. In terms of a general formula,

$$GM = \sqrt[N]{X_1 \times X_2 \times X_3 \times \cdots \times X_N} \qquad \text{(Geometric mean of } N \text{ values)} \qquad (4.10)$$

where GM = geometric mean.
$X_1, X_2, \cdots, X_N$ = series of measurements.
N = number of measurements.

When there are more than two measurements to be averaged in this manner the computations become bothersome, unless we resort to the use of logarithms. The students of mathematics will recognize that if we take logarithms of both sides of formula (4.10) we obtain the equation

$$\log GM = \frac{\Sigma(\log X)}{N} \quad \text{(Logarithmic solution of geometric mean)} \quad (4.11)$$

In other words, the steps called for are as follows:

Step 1. Convert each X into a corresponding log X, by using Table K, Appendix C.
Step 2. Sum the log X values.
Step 3. Divide this sum by N. This result is the logarithm of the geometric mean, as shown by formula (4.11).
Step 4. Find the antilog of the value obtained in step 3. This is the geometric mean.

These steps are illustrated in Table 4.9, which will be explained next.

TABLE 4.9.—COMPUTATION OF A GEOMETRIC MEAN OF TONES MATCHED FOR LOUDNESS TO A STANDARD TONE

(1) Trial	(2) Stimulus (R)	(3) Logarithm of the stimulus (log R)
1	14	1.1461
2	8	0.9031
3	22	1.3424
4	7	0.8451
5	10	1.0000
Sums.........	61	5.2367
Means.......	12.2	1.0473

Geometric mean (antilog of 1.0473) = 11.2

One of the instances in which the geometric mean applies in psychology is in the averaging of stimulus values in psychophysics, when those stimulus values are used to indicate psychological quantities rather than physical quantities. The data in Table 4.9 are fictitious and were invented to illustrate a point. Let us suppose that an observer with very poor discriminative power was asked to control a sound-generating instrument so

as to produce a sound matching in loudness a tone that he has just previously heard. On five different trials the readings of his settings might be as given in column (2) of Table 4.9. We want to find his average setting. The arithmetic mean, as shown in column (2), would be 12.2 units. According to what we know about psychophysical relationships this would be incorrect. We are really interested in the mean of his sensory *responses;* the loudness of the tones that he hears. We assume these to lie on a psychological scale whereas the stimuli lie on a scale of physical energy. Let a value on the psychological scale be called S and one on the physical scale be called R. From Fechner's psychophysical law, the relationship of S to R is usually stated in the equation $S = C(\log R)$. Strictly speaking, the R values should be expressed as multiples of the stimulus limen, but that need not concern us particularly here. We may assume that the R values in column (2) are multiples of the threshold stimulus. In this connection the reader may be reminded of the decibel scale for loudness of sounds. The decibel-scale values are proportional to the logarithm of the stimulus. Ten decibels represents a stimulus 10 times as strong physically as the threshold stimulus; 20 decibels one 100 times as strong; 30 decibels 1,000 times, and so on. The physical values increase in a geometric series while the psychological values are assumed to progress in a parallel arithmetical series.

To return to Table 4.9, the logarithms of R are found in column (3). Their sum is 5.2367 and their mean is 1.0473. The antilog of this value is 11.2, which is the geometric mean. It will be seen that this value is 1.0 unit smaller than the arithmetic mean of the same stimulus values. We would conclude that for this observer the stimulus that for him seems most equivalent to the standard sound is one of 11.2 units.

When to Use the Geometric Mean.—Probably the most common use of the geometric mean in psychology has already been illustrated, namely, in psychophysics.[1] There are other places in which it may well be preferred, for example, in many instances in which time measurements are used, including reaction-time measurements. The need for a geometric mean may be indicated when distributions are distinctly positively skewed. It is best, however, to look for some rational basis, such as the existence of geometric series, before deciding to compute this kind of mean. A rate-of-growth measurement, for example, often involves a geometric series. An important limitation to mention is that a geometric mean cannot be computed when any measurement in the distribution is zero or negative.

Harmonic Mean.—Like the geometric mean, the harmonic mean is needed because the measurements were not made on an appropriate scale.

[1] See Guilford, J. P., Psychometric methods. New York: McGraw-Hill, 1936.

A common application for it in psychology is in connection with "work-limit" tests. In such tests the score is the amount of time required to complete a fixed quantity of work. The frequency distribution of such scores is often positively skewed. Such tests, if given in the more usual form of "time-limit" tests, would yield scores in terms of units of work accomplished in fixed time. The frequency distributions of such scores are more commonly nearly symmetrical. If the ability or abilities measured are assumed to be normally, or at least symmetrically, distributed in the population from which the sample came, then it is reasonable that the work score is a more representative one than the time score; representative in the sense that it spaces individuals better along a scale of equal units.

The *harmonic mean* (HM) is defined as the reciprocal of the mean of the reciprocals of the measurements. The formula is

$$\frac{1}{HM} = \frac{1}{N}\left(\sum \frac{1}{X}\right) \qquad \text{(Equation defining a harmonic mean)} \qquad (4.12)$$

where N and X are as usually defined.
Taking reciprocals of both sides, we have

$$HM = \frac{1}{\frac{1}{N}\left(\sum \frac{1}{X}\right)} \qquad \text{(Same as (4.12) in reciprocal form)} \qquad (4.13)$$

$$HM = \frac{N}{\sum \frac{1}{X}} \qquad \text{(Computing formula for harmonic mean)} \qquad (4.14)$$

The computational steps are as follows:

Step 1. Convert every measurement into its reciprocal. For the purposes of this method, at least three significant figures should be retained.
Step 2. Sum the reciprocals.
Step 3. Divide N by the sum of the reciprocals.

The computation of HM is illustrated in Table 4.10. The scores are in terms of number of minutes required to complete a series of 180 reactions. They have been invented for the illustration. The reciprocals are found in column (3). Their sum is .2834 and their mean is .0567. The reciprocal of this number, 17.6, is the harmonic mean. This is interpreted as the average amount of time required to complete the task.

The whole procedure may be made more reasonable by showing that the conversion of each time score into a reciprocal is equivalent to converting it into a work score. In column (4) are shown the "rate" scores in terms of reactions per minute. Person A took 36 min. to complete 180

TABLE 4.10.—COMPUTATION OF A HARMONIC MEAN OF WORK-LIMIT SCORES

(1) Person	(2) Time scores in minutes* (X)	(3) Reciprocals of time scores ($1/X$)	(4) Rate scores in reactions per minute ($180/X$)
A	36	0.0278	5
B	20	0.0500	9
C	18	0.0556	10
D	15	0.0667	12
E	12	0.0833	15
Sums......	101	0.2834	51
Means.....	20.2	0.0567†	10.2‡

* Number of minutes required to complete 180 reactions.
† The reciprocal of .0567 is 17.6, the harmonic mean.
‡ When converted to a time score, this becomes 180/10.2 = 17.6.

reactions. His rate is 5 reactions per minute. And so on, for the others. A little study of the values in column (4) will show that each is 180 times the corresponding value in column (3). The arithmetic mean of the rate scores is 10.2 reactions per minute. This can be converted back to a time score by the ratio 180/10.2 or it might be left as a work-rate score. If it is converted, the result checks with the harmonic mean.

As in the case of the geometric mean, the harmonic mean cannot be computed when any X is zero or negative.

Exercises

DATA 4*A*.—SCORES IN AN ENGLISH-USAGE EXAMINATION

Scores	*f*
52–53	1
50–51	0
48–49	5
46–47	10
44–45	9
42–43	14
40–41	7
38–39	8
36–37	6
34–35	5
32–33	3
Sum..........	68

DATA 4*B*.—AFFECTIVITY SCORES
(Per cent of 400 words marked "pleasant")

Scores	*f*
95–99	6
90–94	11
85–89	16
80–84	7
75–79	9
70–74	8
65–69	2
60–64	3
55–59	2
50–54	1
Sum..........	65

DATA 4C.—SCORES MADE BY GRADUATES AND ELIMINEES IN THE COMPLEX COORDINATION TEST BY STUDENT PILOTS

Scores	Frequencies	
	Graduates	Eliminees
95–99	1	
90–94	1	
85–89	7	1
80–84	13	2
75–79	37	6
70–74	75	23
65–69	189	34
60–64	297	94
55–59	406	144
50–54	425	208
45–49	341	209
40–44	174	205
35–39	81	105
30–34	16	34
25–29	5	15
20–24	0	2
15–19	1	

DATA 4D.—SCORES IN AN ADJUSTMENT INVENTORY OBTAINED FROM ALCOHOLICS AND NONALCOHOLICS OF BOTH SEXES*

Scores	Frequencies			
	Males		Females	
	Alcoholics	Nonalcoholics	Alcoholics	Nonalcoholics
66–71	1			
60–65	6		3	
54–59	13	1	2	1
48–53	13	1	10	2
42–47	17	3	11	1
36–41	33	3	12	1
30–35	32	2	8	8
24–29	32	9	11	17
18–23	23	16	5	26
12–17	24	36	2	40
6–11	7	43	2	49
0–5	1	25		21

* Manson, M. P. A psychometric differentiation between alcoholics and non-alcoholics, *Quar. J. Stud. Alcohol*, 1948, 9, 175–206.

1. Compute the arithmetic mean of any or all distributions in Data 4*A* to 4*F* inclusive, using the method that seems most feasible. In Data 4*E*, you will need to make some assumption about the cases in the two highest intervals. State your assumptions if means are computed for these distributions.

2. Compute medians for any or all distributions in Data 4*A* to 4*F* inclusive. Why is the difficulty experienced with computation of the mean in Data 4*E* not also encountered in computing the median?

3. Give the crude modes for all distributions in Data 4*A* to 4*F*. Compute the estimated mode in distributions for which you know both mean and median.

DATA 4G.—SOME UNGROUPED DATA

a. 8, 15, 13, 6, 10, 16, 7, 12, 11, 14, 9
b. 12, 10, 18, 13, 4, 8, 17, 15, 6, 14
c. 9, 8, 9, 15, 3, 9, 11, 9, 13
d. 12, 28, 19, 15, 15, 35, 14, 15
e. 7, 18, 20, 14, 27, 23, 13, 3

4. Compute and list the means, medians, and crude modes (where possible) for the distributions in Data 4*G*.

DATA 4*E*.—AGES OF COLLEGE FRESHMEN

Age at last birthday	Men	Women
31–35	1	2
26–30	3	6
25	7	6
24	6	7
23	11	7
22	20	6
21	23	16
20	40	13
19	88	48
18	117	67
17	69	57
16	2	6
Sums.........	387	241

DATA 4*F*.—AIMING-TEST SCORES
(In terms of average error in millimeters)

Score	Men	Women
8.0–8.4	1	
7.5–7.9	5	
7.0–7.4	2	
6.5–6.9	7	2
6.0–6.4	6	4
5.5–5.9	11	3
5.0–5.4	10	9
4.5–4.9	16	7
4.0–4.4	18	15
3.5–3.9	19	12
3.0–3.4	17	15
2.5–2.9	17	13
2.0–2.4	14	14
1.5–1.9	13	10
1.0–1.4	8	1
0.5–0.9	1	
Sums.........	165	105

5. For each distribution in Data 4*G*, tell to which measure of central tendency you give first preference and to which, second. Give reasons.

6. For each distribution in Data 4*A* to 4*F* inclusive, tell which measure of central tendency you would prefer and which would be your second choice. Give reasons.

7. Find the weighted mean of the four means: 15, 16, 18, and 21. These means were derived from samples in which the N's were 6, 10, 25, and 20, respectively. Compute the unweighted arithmetic mean of the four, for comparison. Interpret your result.

8. Find the weighted mean of the proportions .25, .30, .32, and .33. These proportions were based upon samples whose N's are 44, 32, 18, and 25, respectively. Compute an unweighted arithmetic mean of these proportions, for comparison. Interpret your results.

9. Find the geometric mean of the number 2, 9, 15, and 16. Compute the arithmetic mean, for comparison. Interpret your results.

10. Find the harmonic mean of the work-limit scores 20, 25, 40, and 50. These scores represent the total time summated in a series of 120 simple reaction times and are in terms of seconds. Interpret your results.

CHAPTER 5

MEASURES OF VARIABILITY

Knowing the central tendency of a set of measurements tells us much, but it does not by any means give us the total picture of the sample we have measured. Two groups of six-year-old children may have the same average *IQ* of 105, from which we would conclude that, taken as a whole, each group is as bright as the other, and we might expect from the two the same average level of performance in school or out of school in areas of life where *IQ* is important. Yet when we are told, in addition, that one group has no individuals with *IQ*'s below 95 or above 115, whereas the other has

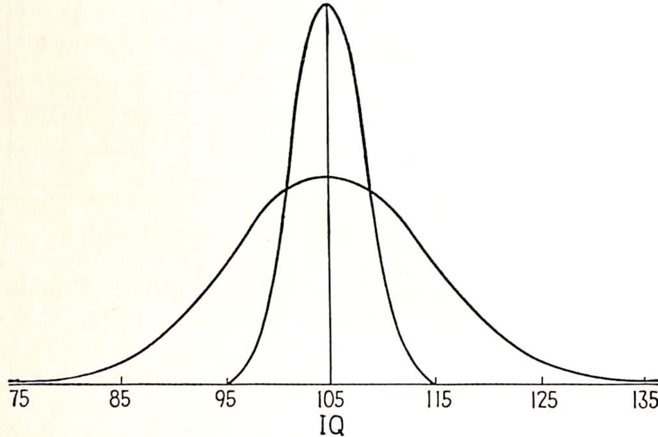


Fig. 5.1.—Two distributions with the same mean (*IQ* = 105) but with decidedly different ranges (and dispersions).

individuals with *IQ*'s ranging from 75 to 135, we recognize immediately that there is a decided difference between the two groups in variability or dispersion of brightness. The first group is decidedly more homogeneous with respect to *IQ*, and the second is decidedly more heterogeneous. We should expect the first group to be much more teachable in that they will grasp new ideas at about the same rate and progress at about the same rate. We should expect the second group to show considerable disparity in speed of grasping new ideas. There will be extreme laggards at the one end of the distribution and others at the other end of the distribution who may be irked at the slow progress of the group. The distributions for two such groups, when plotted, resemble those in Fig. 5.1.

It is the purpose of this chapter to explain and illustrate the methods of indicating degree of variability or dispersion by the use of single numbers, just as in the preceding chapter we saw how the central tendency of a distribution could be indicated by a single number. The four most customary values to indicate variability are (1) the total range, (2) the semi-interquartile range Q, (3) the standard deviation σ, and (4) the average (or mean) deviation AD.[1]

The Total Range

The total range is the indicator of variability that is easiest and most quickly ascertained but is also the most unreliable; thus it is almost entirely limited to the purpose of preliminary inspection. In the illustration of the preceding paragraph, the range of the first group (from an *IQ* of 95 to an *IQ* of 115) was 21 *IQ* points inclusive. The range of the second group was from 75 to 135 *IQ* points. The range is the distance given by highest score minus lowest score, plus 1. From this comparison, we draw the conclusion that the second group is considerably more variable than the first.

Why the Range Is Unreliable.—The range is very unreliable for the reason that only two measurements are used to determine it. The remaining measurements have nothing to do with the estimation of it. In the second group just mentioned, it might have been true that there were several *IQ*'s of 75 and also several *IQ*'s of 135; but this would be most unusual. The chances are great that there would be only one 75 and one 135. Furthermore, the next lowest *IQ* might have been 85, with a gap of 10 points to the very lowest; and the next to the highest might have been 120, a distance of 15 points from the very highest. Had either or both of the persons with 75 *IQ* and 135 *IQ* been missing from the group, the range would have been something very different from the 61 points actually obtained. This is what we mean by saying that the total range is highly unreliable. Some faith can, of course, be placed in it when there is more than one case having each of the extreme measurements and when there are no decided gaps in the tails of the distribution.

When Ranges Should Not Be Compared.—Total ranges should not be compared when two distributions have a markedly different number of cases. It is quite natural for more extreme cases to show up as we add new cases to any sample, so that larger groups should be expected to have wider total scatter. This factor is not nearly so important for other indi-

[1] The *probable error PE* has been used as a measure of variability, but it seems rapidly to be going out of use and so is merely mentioned in this volume (see the footnote on p. 104).

cators of dispersion as it is for total range. Another caution almost goes without saying, and that is the impossibility of comparing ranges in two distributions where the units of measurement are not the same.

The Semi-interquartile Range—Q

The semi-interquartile range, Q, is *one-half* the range of the middle 50 per cent of the cases. First we find by interpolation the range of the middle 50 per cent, or interquartile range, then divide this range by 2. See Fig. 5.2 for a general picture of the relation of Q to a frequency distribution.

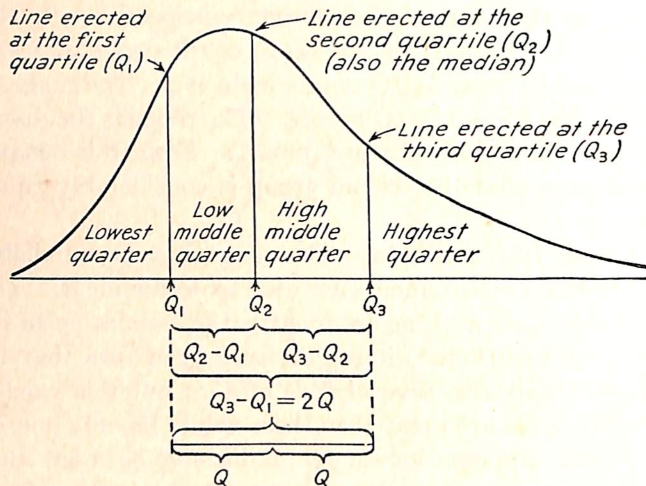


Fig. 5.2.—Illustration of the quartiles Q_1, Q_2, and Q_3, the interquartile and semi-interquartile ranges, and the quarters of the sample in a slightly skewed distribution.

Quartiles and Quarters.—When we count up from below to include the lowest or first quarter of the cases, we find the point called the *first quartile*, which is given the symbol Q_1. Counting down from above to include the highest or fourth quarter of the cases, we locate the third quartile, or Q_3. Incidentally, the median, which separates the second and third quarters of the distribution, is also called Q_2. Note that the quartiles Q_1, Q_2, and Q_3 are *points* on the measuring scale. They are division points between the *quarters*. We may say of an individual that he is *in* the highest *quarter* (or fourth quarter), and we may say of another that he is *at* the third *quartile*. We should never say of an individual that he is *in* a certain *quartile*.

Interpolation of Q_1 *and* Q_3.—In the distribution of ink-blot scores again, we locate the third and first quartiles by interpolation (see Table 5.1).

TABLE 5.1.—DETERMINATION OF Q_3, Q_1, and Q (THE SEMI-INTERQUARTILE RANGE) FOR THE INK-BLOT TEST SCORES

Scores	*f*	
55–59	1	
50–54	1	
45–49	3	
40–44	4	
35–39	6	← Q_3 lies within this interval
30–34	7	
25–29	12	
20–24	6	← Q_1 lies within this interval
15–19	8	
10–14	2	
	$N = 50$	

$$Q_1 = 19.5 + \frac{2.5}{6} \times 5 = 19.5 + 2.08 = 21.58$$

$$Q_3 = 39.5 - \frac{3.5}{6} \times 5 = 39.5 - 2.92 = 36.58$$

$$Q = \frac{36.58 - 21.58}{2} = \frac{15.00}{2} = 7.5$$

One-fourth of the cases ($N/4$) is 12.5. Counting up from the bottom to include 12.5 cases, we find that we need 2.5 out of the 6 cases in the third class interval. As in earlier solutions, 2.5/6 times 5 gives 2.08. Added to 19.5, this gives 21.58 as the position of Q_1. Counting down from the top, we find that we need 3.5 cases out of 6 in the fifth class interval. Then 3.5/6 of 5 gives 2.92. Deducted from 39.5, this leaves 36.58 as our estimate of Q_3.

The Interquartile Range and Q.—The interquartile range, or the distance from Q_1 to Q_3, is given by $Q_3 - Q_1$, or 36.58 − 21.58, which equals 15.00. The semi-interquartile range is one-half of this, or 7.5. In terms of a formula,

$$Q = \frac{Q_3 - Q_1}{2} \quad \text{(Semi-interquartile range)} \quad (5.1)$$

where Q_3 = third quartile.
Q_1 = first quartile.

How Quartiles Indicate Skewness.—It is of interest in passing to take note of the relative distances of Q_3 and Q_1 from the median, or Q_2, in a distribution. If the distribution is exactly symmetrical, both the third and first quartiles will be the same distance from the median, and that distance is Q. When there is any skewness in the distribution, the two

distances will be unequal. If the skewness is positive, the distance $Q_3 - Q_2$ will be greater than the distance $Q_2 - Q_1$. If the skewness is negative, the reverse will be true. In other words, skewness is:

positive when $(Q_3 - Q_2) > (Q_2 - Q_1)$
negative when $(Q_3 - Q_2) < (Q_2 - Q_1)$
and zero when $(Q_3 - Q_2) = (Q_2 - Q_1)$

The relative sizes of these two distances therefore tells much about the direction and the amount of skewness in the distribution. For the ink-blot scores, $Q_3 - Q_2$ is 8.4, and $Q_2 - Q_1$ is 6.6. Our inference is that the distribution is positively skewed to a moderate degree. In Fig. 5.2 the distribution is positively skewed and $(Q_3 - Q_2)$ is greater than $(Q_2 - Q_1)$.

The Average Deviation

The average deviation, or AD, is the arithmetic mean of all the deviations when we disregard the algebraic signs. Every score or measurement in a distribution deviates from the mean in that it is a certain distance above or below the mean. When and if any measurement coincides exactly with the mean, its deviation is zero. Deviations above the mean are regarded as positive distances; those below the mean as negative distances. In terms of an algebraic definition,

$$x = X - M \qquad \text{(A deviation of a measurement from the mean)} \qquad (5.2)$$

where X = an original score or measurement.
M = the arithmetic mean.

As was pointed out in a previous chapter, the deviations from the mean may be regarded as *moments* about a center of gravity. If we sum the deviations, taking into account the algebraic signs, the sum would be zero. In other words, $\Sigma x = 0$. The average of the deviations would also be zero, because $\Sigma x/N = 0/N$, and zero divided by any finite number is equal to zero. This kind of an average of the deviations tells us nothing, therefore, about their size. We want some indication of their over-all size in order to describe the amount of dispersion. The greater the spread of the deviations, the greater the dispersion of the distribution. One solution is to disregard the algebraic signs of the deviations. In doing so, we disregard their direction; we are interested only in their amount. We treat them as if they were all positive. In terms of a formula,

$$AD = \frac{\Sigma|x|}{N} \qquad \text{(The average deviation)} \qquad (5.3)$$

where $|x|$ (with the vertical bars embracing it) = an absolute value of x, *i.e.*, disregarding algebraic sign.

TABLE 5.2.—CALCULATION OF THE AVERAGE DEVIATION IN UNGROUPED DATA
(Mean = 13.2)

X	$\|x\|$
13	0.2
17	3.8
15	1.8
11	2.2
13	0.2
11	2.2
17	3.8
13	0.2
11	2.2
11	2.2
	18.8
	$\Sigma\|x\|$

$$AD = \frac{18.8}{10} = 1.88, \text{ or } 1.9$$

To illustrate the solution of an average deviation, consider Table 5.2. The sum of the absolute deviations is 18.8. Divided by N, this gives 1.88 as the average deviation. Because of the small size of N, we should round to one decimal place and give the AD as 1.9.

Interpretation of an Average Deviation.—From the formula and the computations it will be seen that when we compute the average deviation we are interested merely in the *size* of the deviations from the mean. We ignore their direction. The AD is an arithmetic mean of all the deviations of whatever size or direction. Like any arithmetic mean, it stands for all the values averaged. In the problem just solved, the AD tells how much on the average the different observations of the auditory limen differed from their mean, 13.2. The answer is that on the average these deviations were 1.9 cycles, or a little less than 2.

In samples that are not too small and when distributions approach the normal bell-shaped form, we may make the further remark that about 58 per cent of the observations should be expected to fall within the limits 1 AD below the mean and 1 AD above the mean. In the threshold problem those two conditions are not satisfied; the distribution is neither large enough nor symmetrical enough to warrant such a conclusion. If this were the case, however, we could say that 58 per cent of the 10 measurements (6 of them) should be expected between 13.2 − 1.9 = 11.3 and 13.2 + 1.9 = 15.1. This would include all integral values of 12, 13, 14,

and 15. Actually, only four of the observations were included within those limits, though this should not surprise us, in view of the smallness of the sample.

Computation of the *AD* from Grouped Data.—Although the average deviation is not often computed for large, regular samples in ordinary statistical practice, it is probably worth demonstrating how this statistic can be conveniently computed from data grouped in class intervals. Table 5.3 demonstrates this kind of solution. The mean of the 50 ink-blot

TABLE 5.3.—COMPUTATION OF AN AVERAGE DEVIATION IN GROUPED DATA

(1) Scores	(2) X	(3) x	(4) f	(5) fx
55–59	57	+27.4	1	+ 27.4
50–54	52	+22.4	1	+ 22.4
45–49	47	+17.4	3	+ 52.2
40–44	42	+12.4	4	+ 49.6
35–39	37	+ 7.4	6	+ 44.4
30–34	32	+ 2.4	7	+ 16.8
25–29	27	− 2.6	12	− 31.2
20–24	22	− 7.6	6	− 45.6
15–19	17	−12.6	8	−100.8
10–14	12	−17.6	2	− 35.2
Sums.....			50	425.6
			N	$\Sigma\|fx\|$

test scores represented in Table 5.3 was previously reported as 29.60. Ordinarily, one decimal place (or one digit beyond the last at the right in the original measurements) will do in the computation of the *AD*.

Column (2) of Table 5.3 presents the midpoints of the intervals. The midpoint value represents every measurement in the interval. Column (3) gives the deviations of these midpoints from the computed mean. Algebraic signs are recorded for the sake of accuracy but they will not be needed in the computations. In column (5) are the products of each frequency times its corresponding deviation, in other words, each fx product. The equation for the *AD* by this procedure is

$$AD = \frac{\Sigma|fx|}{N} \qquad \text{(The average deviation from grouped data)} \qquad (5.4)$$

where f, x, and N are as previously defined, and the fx products are summed without regard to algebraic sign. From the data in Table 5.3,

$$AD = \frac{425.6}{50}$$
$$= 8.512$$

which should be rounded to 8.5.[1]

According to the kind of interpretation given previously, we may say that if this distribution of scores is close to normal, we should expect 58 per cent of the scores to lie between 21.1 and 38.1. This would mean 29 of the 50 scores. Since the data are grouped in Table 5.3, we cannot check this conclusion by actual count of the cases, but a rough check can nevertheless be made. If we assume that the 6 individuals in the interval 35–39 are evenly distributed, about 4 of them should be below 38.1. If we assume, likewise, that the 6 individuals in the interval 20–24 are evenly distributed, then 4 of them should be above the point 21.1. With these assumptions made, there are 27 cases between the points 21.1 and 38.1. This number is 54 per cent of the sample. Fifty-eight per cent would have called for 29. The agreement may be regarded as close enough, in view of the fact that the sample is not so very large and the fact that it tends to be positively skewed. Such a check is often sufficient to tell us whether we have made any *serious* errors in computing the average deviation by this method.

The Standard Deviation

The standard deviation, or σ, is the most commonly used indicator of degree of variability, and of the ones described in this chapter it is usually the most reliable. That is, it varies least from sample to sample drawn at random from the same population. It is therefore more dependable and, as an estimate of the dispersion of the population, it is more accurate.

Computing the Standard Deviation Directly from Deviations.—Like the *AD*, the standard deviation is also a kind of average of all the deviations about the mean in a sample, though it is not a simple arithmetic mean.[2] The fundamental formula for it is

$$\sigma = \sqrt{\frac{\Sigma x^2}{N}} \quad \text{(Basic formula for the standard deviation in a sample)} \quad (5.5)$$

[1] One check on the accuracy of computations of the Σfx values is to sum them algebraically. The sum Σfx should equal approximately zero, small discrepancies due to rounding errors being tolerated. In Table 5.3, Σfx equals exactly zero.

[2] In some textbooks the standard deviation of a sample is symbolized by the double lettering *SD*, or *S.D.* In some others it is denoted by the letter *s*. The lack of agreement is unfortunate. The author believes that the symbols used here are most common to psychological and educational literature and hence will be better understood by readers in those areas.

where x = deviation from the mean of the sample.
N = the size of the sample.

Some Fundamental Concepts: Sum of Squares; Variance.—Formula (5.5) deserves close study. It calls for several steps in fixed order:

Step 1. Find each deviation from the mean (x).
Step 2. Square each deviation, finding x^2.
Step 3. Sum the squared deviations, finding Σx^2.
Step 4. Divide this sum by N, finding $\Sigma x^2/N$.
Step 5. Extract the square root of the result of step 4. This is the standard deviation.[1]

In verbal terms, a standard deviation is the square root of the arithmetic mean of the squared deviations of measurements from their mean. It has often been called the *root-mean-square deviation.* But in this simplified statement lies considerable meaning. Latent in the few steps enumerated above lie two statistical concepts that have increasing importance. One is the *sum of squares*, the end result of step three. The other is called *variance*, the end result of step four. These ideas are best introduced by means of an illustration.

TABLE 5.4.—DATA ILLUSTRATING SUM OF SQUARES, VARIANCE, AND STANDARD DEVIATION

(1) Person	(2) Score X	(3) Deviation x	(4) Deviation squared x^2
A	15	+5	25
B	14	+4	16
C	11	+1	1
D	10	0	0
E	9	−1	1
F	7	−3	9
G	4	−6	36
Sums...............	$70 = \Sigma X$	$0 = \Sigma x$	$88 = \Sigma x^2$
Means...............	10.0	0.0	$12.57 = V$
Standard deviation....	....	....	$3.55 = \sigma$

In Table 5.4 are listed seven fictitious scores representing a sample of seven individuals A to G inclusive. These are denoted by the usual symbol, X. The mean of these seven scores, as shown in column (2), is exactly 10.0. Column (3) shows the deviations of these scores from

[1] These steps are illustrated in Tables 5.4 and 5.5, and in Fig. 5.3.

the mean. Their sum is zero and also their mean, as is to be expected. In column (4) we find the squared deviations. Their sum, 88, is the *sum of squares*. Their mean is equal to 12.57, which we have defined as the *variance*, in this sample. The square root of this is 3.55, the standard deviation. All this follows from formula (5.5) and from the steps and definitions given above. Let us see what this means in terms of a geometrical view of the problem.

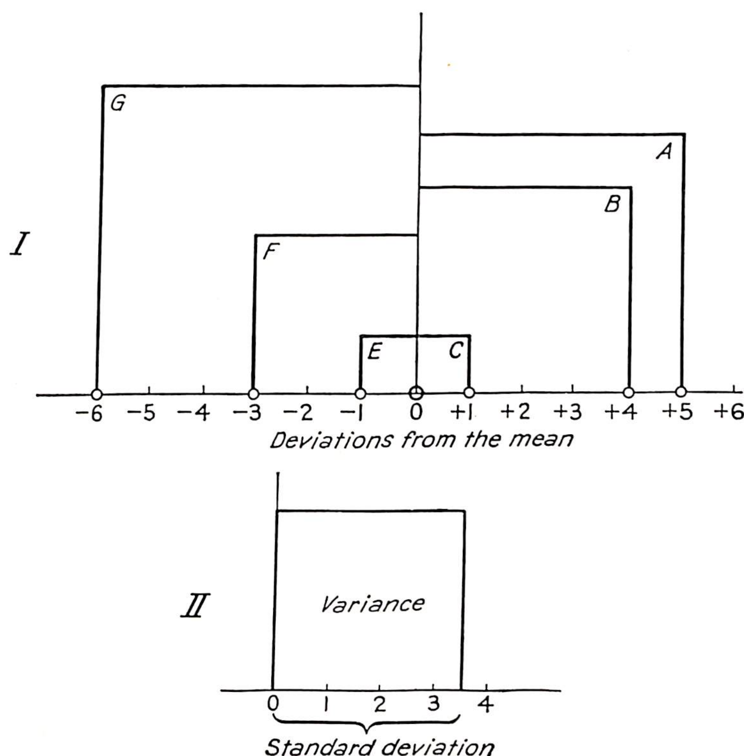


Fig. 5.3.—Illustration of deviations from the arithmetic mean, their squares, their mean (which is the variance), and the standard deviation (which measures the variability) in a sample of seven cases.

For a geometrical representation of these ideas, see Fig. 5.3. In the first diagram, the scale of measurement is shown, as usual, in the form of a straight line extending from left to right. Here, however, the original score values are not marked. The mean has become recognized as the main reference point and has been called zero. This is what happens when we derive deviations x from original scores X. All seven individuals still retain their relative positions, in correct rank order and at the same separations, as they had before. We have merely moved the zero point 10 units up the linear scale.

So much for representing deviations. It will be seen that the points on the line correspond exactly with the values in column (3) of Table 5.4. Consider now the squaring of the deviations. Where deviations themselves are represented by linear distances from a common reference point, squared deviations must be represented by areas, namely, squares. The squares belonging to the different individuals A to G are shown in Fig. 5.3. The areas of the squares are equal numerically to the values given in column (4) of Table 5.4. It can be seen that the individuals come in the same rank order when we compare the squared deviations as when we compare x distances. It is also notable how large deviations, when squared, increase much more relatively than do small deviations. This point will be important to consider later.

The sum of the squares would be represented geometrically as an area equal to a composite of all the squares in Fig. 5.3 I. This could also be shown as a square or as a rectangle. Its dimensions could vary somewhat but its surface would contain 88 units such as those representing persons C and E. Finding the arithmetic mean of this large area is equivalent to apportioning it equally among the seven individuals. It is the amount of area that each person would possess if each one of them were given the same amount. This is the variance, which we may represent in the form of a square in Fig. 5.3 II. This square is shown on a base line like that in the first diagram. Its length of side is the square root of its area and represents the standard deviation.

Some important algebraic relationships, latent in formula (5.5), may be called to the attention of the reader. They are all important for general orientation in this topic. They may be useful not only in thinking about the concepts of sums of squares, variances, and standard deviations, but will be found to enter into computations of various kinds later. First, one more symbol needs to be introduced. *V is often used to stand for variance.* With this additional symbol given, we can state the following interrelationships:

$$\sigma = \sqrt{\frac{\Sigma x^2}{N}} = \sqrt{V} \qquad (5.6)$$

$$V = \frac{\Sigma x^2}{N} = \sigma^2 \qquad \text{(Interrelationships of } \Sigma x^2,\ V, \text{ and } \sigma) \qquad (5.7)$$

$$\Sigma x^2 = NV = N\sigma^2 \qquad (5.8)$$

Both V and σ, each in its own way, are indicators of amount of dispersion in a distribution. V is said to measure variance, σ to measure variability. When the sample is one of individuals measured on a common scale, either V or σ can become familiar indicators of the extent of

the individual differences. To make these concepts more meaningful, then, it is well to think of them in terms of measures of individual differences.

Suppose, first, that we have a sample of only one case; with only one score. There is no possible basis for individual differences in such a sample, and therefore there is no variance or variability. Bring into the picture a second individual with his score in the same test or experiment. We now have one difference. Bring in a third case and we then have two additional differences; three altogether. Bring in a fourth, a fifth, and so on. There are as many differences as there are possible pairs of individuals. We could compute *all* these interpair differences and could average them to get a single, representative value. We could also square them and then average them. It is far more economical, however, to find a mean of all the scores and to use that value as a common reference point. Each difference then becomes a deviation from that reference point and there are only as many deviations as there are individuals. Either the variance or the standard deviation is a single representative value for all the individual differences when taken from a common reference point.

Consider the matter from a somewhat different point of view. Consider giving a certain test of n items to a group of persons. Before giving the first item to the group, so far as this test is concerned the individuals are all alike. All have scores of zero. There is no variance. This may seem absurd, but it has a very reasonable bearing on what comes next. Next administer the first item in the test to all individuals in the group. Some will pass it and some will fail. Some will now have scores of 1 and some still have scores of zero. There are two groups of individuals. There is this much differentiation; this much variance. Give a second item. Of those who passed the first, some will pass the second and some will fail it, unless the two items are perfectly correlated. Of those who failed the first, some may pass the second and some may fail it. There are now three possible scores, 0, 1, and 2. More variance has been introduced. Carry the illustration further, adding item by item. The differences among scores will keep increasing, and so, by computation, also the variance and the variability, as indicated by V and by σ. Psychological and educational testing depend almost entirely upon the phenomenon of individual differences and therefore upon variance. Probably less than one per cent of the tests commonly used yield scores on an absolute scale. The significance of any score is ordinarily its usefulness in placement of a person somewhere in the group. The greater the variance among the scores, the more accurately (usually) each person is placed. Thus, in addition to the use of the standard deviation in describing the spread or

scatter of a certain sample, there is its use, as we shall see in later chapters, in the evaluation of tests and test items in a number of ways (see Ch. 17). After this digression, let us return to the descriptive use of σ and its computation in a typical laboratory problem.

As an illustrative problem in computing σ by formula (5.5), let us take the 10 measurements of the threshold for pitch (see Table 5.5). Their

TABLE 5.5.—CALCULATION OF THE STANDARD DEVIATION IN UNGROUPED DATA

(1) X Scores	(2) x Deviations	(3) x^2
13	−0.2	.04
17	+3.8	14.44
15	+1.8	3.24
11	−2.2	4.84
13	−0.2	.04
17	+3.8	14.44
13	−0.2	.04
11	−2.2	4.84
11	−2.2	4.84
11	−2.2	4.84
		51.60 Σx^2

$$\sigma = \sqrt{\frac{51.60}{10}} = \sqrt{5.160} = 2.27, \text{ or } 2.3$$

mean we found to be 13.2. The deviations from the mean are given in column (2) and their squares, in column (3). Their sum is 51.60. The mean of the squared deviations is 5.160. The standard deviation is the square root of this, or 2.27. This should not be reported to more than one decimal place. In terms of the unit of our measuring scale, this is 2.3 cycles per second.

The Interpretation of a Standard Deviation.—Now that we have the answer 2.3 cycles per second, how shall we interpret it? The usual and most accepted interpretation is in terms of the percentage of cases included within the range from one standard deviation below the mean to one standard deviation above the mean. This range on the scale of measurement includes about two-thirds of the cases in the distribution. In a normal distribution, it is known that from -1σ (one standard deviation below the mean) to $+1\sigma$ (one standard deviation above), exactly 68.26 per cent of the cases are found. Since most samples yield distributions

that depart to some degree from normality, we say, "about two-thirds," which is, of course, a little short of 68.26 per cent. Fig. 5.4 illustrates the division of the area under a normal curve into regions marked off at -1σ and $+1\sigma$. With two-thirds of the surface *within* those limits, there is left one-third of the area to be divided between the two "tails" of the distribution—one-sixth below the point at -1σ and one-sixth above the point at $+1\sigma$.

In the problem just solved, where we found σ equal to 2.3, the distance from -1σ to $+1\sigma$ on the scale of measurement is 10.9 to 15.5 cycles; *i.e.*, the mean 13.2 minus 2.3 is 10.9, and the mean plus 2.3 is 15.5 cycles. Within these limits are all measurements of 11, 12, 13, 14, and 15. By actual count, there are four 11's, three 13's, and one 15, or 8 of the 10 measurements within these limits, whereas we should have expected 7.

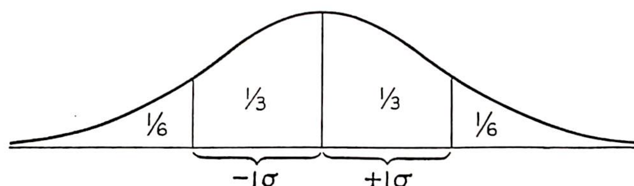


Fig. 5.4.—Approximate fractions of the area under a normal distribution curve (consequently, fractions of the N cases in a normally distributed sample) that lie within one standard deviation of the mean and also beyond the limits of one standard deviation, in either direction.

But, because of the small number of cases and the fact that the distribution is irregular, we should not be surprised at this result. In other problems this comparison serves as a rough check upon the accuracy of computation of σ. It will not catch all errors but will indicate gross errors if the sample is not too small and the distribution is fairly normal.

Grouping Deviations as a Short Cut.—Some saving in time and effort can be afforded in the solution of the standard deviation in data like those in Table 5.5, if we group them as in Table 5.6. Since the same measurement is repeated several times and its deviation from the mean is the same every time, and also its deviation squared, we need to find the deviation and its square only once and multiply each x^2 by its frequency. The last column of Table 5.6 contains the fx^2 products, and it will be seen that their sum is again 51.60, from which the standard deviation will be the same as before. The formula for this reads

$$\sigma = \sqrt{\frac{\Sigma fx^2}{N}} \quad \text{(Standard deviation from grouped data)} \quad (5.9)$$

where the symbols are defined as before.

TABLE 5.6.—CALCULATION OF THE STANDARD DEVIATION IN GROUPED DATA WITH THE USE OF ACTUAL DEVIATIONS

(1) X	(2) x	(3) x^2	(4) f	(5) fx^2
17	+3.8	14.44	2	28.88
15	+1.8	3.24	1	3.24
13	−0.2	.04	3	.12
11	−2.2	4.84	4	19.36
				51.60 Σfx^2

A similar treatment may be given all grouped data, in which we let the midpoint of each interval be the X for all cases within the interval, and this X minus M gives us the deviation of all cases within the interval. From here on, the procedure is the same as that in Table 5.6. We shall not illustrate the steps by means of a special problem, for there are more efficient ways of dealing with grouped data.

The Standard Deviation by the Short Method.—The short method, which was employed in the preceding chapter to calculate a mean (Table 4.2), will now be extended in order to compute a standard deviation. The first steps are carried out exactly as previously to the point of finding the mean. The mean itself need not be known (since we are dealing with a guessed mean), but the correction is required, as will be seen in the following formula:[1]

$$\sigma = i\sqrt{\frac{\Sigma fx'^2}{N} - c'^2} = i\sqrt{\frac{\Sigma fx'^2}{N} - \left(\frac{\Sigma fx'}{N}\right)^2}$$

(Standard deviation from grouped and coded data) (5.10)

where i = size of class interval.

x' = deviation from the guessed mean in terms of the class interval as the temporary unit.

c' = correction in the guessed mean, also in terms of the class interval as the unit.

For computational convenience the formula may be varied as follows:

$$\sigma = \frac{i}{N}\sqrt{N\sum fx'^2 - \left(\sum fx'\right)^2}$$

(Alternate for formula 5.10) (5.11)

[1] This formula should be better understood by the student who follows the proofs given in Appendix A.

TABLE 5.7.—CALCULATION OF THE STANDARD DEVIATION USING THE SHORT METHOD (GUESSED-MEAN PROCEDURE)

(1) Score	(2) f	(3) x'	(4) fx'	(5) fx'^2
55–59	1	+5	+ 5	25
50–54	1	+4	+ 4	16
45–49	3	+3	+ 9	27
40–44	4	+2	+ 8	16
35–39	6	+1	+ 6	6
30–34	7	0	0	0
25–29	12	−1	−12	12
20–24	6	−2	−12	24
15–19	8	−3	−24	72
10–14	2	−4	−8	32
	50 N		−24 $\Sigma fx'$	230 $\Sigma fx'^2$

$$c' = \frac{\Sigma fx'}{N} = \frac{-24}{50} = -.48$$

$$\sigma = 5\sqrt{{}^{230}\!/_{50} - (-.48)^2} = 5\sqrt{4.6 - .2304} = 5\sqrt{4.3696} = 5 \times 2.09 = 10.45$$

The procedure is illustrated in Table 5.7, which is similar to Table 4.2 through column (4). For all class intervals, we need to know the fx'^2 products, and these are given in column (5). In each row, the fx'^2 product is found by multiplying the corresponding numbers in columns (3) and (4); *i.e.*, the first one, 25, is the product of 5×5; the second one is the product of 4×4; and the third, the product of 3×9; etc. This is because the product fx'^2 may be factored as $(fx')x'$. It is excellent checking procedure to do the multiplying also by the product $(f) \times (x'^2)$ for each interval.

Next we sum the fx'^2 products to obtain $\Sigma fx'^2$. In Table 5.7, this is 230. To find c', we divide $\Sigma fx'$ by N. In this case, it is −24/50, which equals −0.48. We need c'^2, which is 0.2304. Now, to apply formula (5.10), we need next to divide $\Sigma fx'^2$ by N, or 230/50, which equals 4.6. Deduct c'^2 from this, or 4.6 − 0.2304, and we have 4.3696. The square root of this is called for next, and this is 2.09. The last step is to multiply by i, the size of the class interval; 2.09×5 equals 10.45, which is the standard deviation we have been seeking.

We may now say that about two-thirds of the individuals should be expected between the mean minus 10.45 and the mean plus 10.45. Since the mean is 29.6, these limits are 19.2 and 40.0. Fortunately, for the

sake of checking on this conclusion, these limits are close to the division points between class intervals (see Table 5.7). The four intervals included within these limits have in them 31 cases altogether, which are 62 per cent of the whole group. This is a little short of two-thirds but not unreasonably so.[1]

Rough Checks for the Solution of the Standard Deviation.—The kind of comparison just mentioned is a rough check for the correct solution of the standard deviation. If the actual percentage of cases between $+1\sigma$ and -1σ deviates too far from 68 per cent, there is probably something wrong with the calculation, and a recalculation is in order. This check cannot often be satisfactorily applied with grouped data because the frequencies from -1σ to $+1\sigma$ cannot then be accurately determined.

Another rough check is to compare the standard deviation obtained with the total range of measurements. In large samples ($N = 500$ or more) the standard deviation is about one-sixth of the total range. Stated in other terms, the total range is about 6 standard deviations. In smaller samples, the ratio of range to standard deviation becomes smaller, as indicated in Table 5.8.

TABLE 5.8.—RATIOS OF THE TOTAL RANGE TO THE STANDARD DEVIATION IN A DISTRIBUTION FOR DIFFERENT VALUES OF N*

N	Range/σ	N	Range/σ	N	Range/σ
5	2.3	40	4.3	400	5.9
10	3.1	50	4.5	500	6.1
15	3.5	100	5.0	700	6.3
20	3.7	200	5.5	1,000	6.5

* Adapted from Snedecor, G. W. Statistical methods. P. 85. Ames, Iowa: Collegiate, 1940.

In the ink-blot data, since $N = 50$, we should expect the range to be 4.5 times the standard deviation. The standard deviation 10.45 times 4.5 gives us an expected range of about 47 points. Actually the range was 46 points, which checks so closely as to give us confidence that our standard deviation is at least not grossly in error.

[1] The probable error of a distribution is computed directly from the standard deviation by the formula $PE = .6745\sigma$. It is numerically about two-thirds as large as σ, as suggested by the ratio .6745. One more multiplication is required; consequently it is not quite so easily computed as the standard deviation. Its chief virtue is that *in a normal distribution*, 50 per cent of the measurements lie between the points at $-1PE$ and $+1PE$ from the mean. The writer has come to feel that this is not sufficient excuse for the inclusion of one more statistic to the already lengthy list, particularly when the actual middle 50 per cent of the measurements can be more certainly delimited by the interval $Q_3 - Q_1$, the use of which does not assume normality of distribution.

It may seem strange that we use a less reliable statistic like range as a criterion of accuracy of a more reliable statistic like the standard deviation. The reasons are that (1) there can hardly be any error in computing such a simple thing as the range, whereas (2) there are chances of gross errors in calculating σ because of the many steps involved, for example, failing to make the final step of multiplying by i.

A Summary of Steps for Computing the Standard Deviation.—The steps necessary for the calculation of σ by the short method are as follows:

Step 1. Complete Steps 1 through 6 already listed for finding the mean by the guessed-average route (see Table 4.2).
Step 2. Find for every class interval the fx'^2 product. The most efficient way is to compute the product of x' times fx' for each interval. These products will all be positive.
Step 3. Sum the fx'^2 products.
Step 4. Divide this sum by N, carrying to at least two decimal places.
Step 5. Find c'^2, to at least two decimal places.
Step 6. Deduct the number found in Step 5 from that found in Step 4.
Step 7. Find the square root of the number found in Step 6, keeping two decimal places.[1]
Step 8. Multiply this number by the size of the class interval. If N is large, report two decimal places; if small, round to one decimal place.
Step 9. Interpret the standard deviation in terms of the two-thirds principle.
Step 10. Apply the rough check of comparing σ with the range and using the ratios of Table 5.8.

The Standard Deviation from Original Measurements.—If the number of measurements is not large, if the measurements themselves are small numbers, particularly when a good calculating machine is available, the best procedure for computing a standard deviation is by means of the formula

$$\sigma = \frac{1}{N}\sqrt{N\sum X^2 - \left(\sum X\right)^2} \qquad \text{(Standard deviation computed without knowledge of deviations)} \quad (5.12)$$

in which the essential steps are:

Step 1. Square each score or measurement.
Step 2. Sum the squared measurements to give ΣX^2.

[1] In this, and in the following steps, it is assumed that we are dealing with integral measurements. If they are in terms of decimal fractions or multiples of 10 or 100, this rule applies only after making the necessary allowance for the place of the decimal point.

Step 3. Multiply ΣX^2 by N to give $N\Sigma X^2$.
Step 4. Sum the X's to find ΣX.
Step 5. Square the ΣX to find $(\Sigma X)^2$.
Step 6. Find the difference $N\Sigma X - (\Sigma X)^2$.
Step 7. Find the square root of the number found in Step 6.
Step 8. Divide the number found in Step 7 by N (or multiply it by $1/N$).

On the calculating machine, the X's and the X^2's can be accumulated at the same time according to instructions provided with the machine. In tabular form, the solution of this kind is illustrated in Table 5.9.

TABLE 5.9.—CALCULATION OF THE STANDARD DEVIATION FROM THE ORIGINAL MEASUREMENTS AND UNGROUPED DATA

X	X^2
13	169
17	289
15	225
11	121
13	169
17	289
11	121
13	169
11	121
11	121
132	1,794
ΣX	ΣX^2

$$\sigma = \tfrac{1}{10}\sqrt{10(1{,}794) - 132^2}$$
$$= \tfrac{1}{10}\sqrt{17{,}940 - 17{,}424}$$
$$= \tfrac{1}{10}\sqrt{516}$$
$$= \frac{22.7}{10}$$
$$= 2.27, \text{ or } 2.3$$

Grouping Original Measurements.—If the scores are conveniently grouped and their frequencies tabulated, as in Table 5.10, some saving

TABLE 5.10.—CALCULATION OF THE STANDARD DEVIATION FROM THE ORIGINAL MEASUREMENTS, WITH GROUPING

X	f	fX	X^2	fX^2
17	2	34	289	578
15	1	15	225	225
13	3	39	169	507
11	4	44	121	484
	10	132		1,794
	N	ΣfX		ΣfX^2

in work can be effected. The steps by which we arrive at ΣfX and ΣfX^2 should now be easy to follow by an analogy to the last previous solution. Once those values are obtained, Steps 6 to 8 above can be followed to arrive at σ. The formula for this procedure is

$$\sigma = \frac{1}{N}\sqrt{N\sum fX^2 - \left(\sum fX\right)^2} \quad \text{(Same as formula 5.12, with grouped data)} \qquad (5.13)$$

Correction of the Standard Deviation for Coarse Grouping.—We are now ready to see more clearly why the number of class intervals should not be too small in grouping data or the class interval too large. Reference was previously made (p. 61) to a "grouping error." Let us see what the grouping error is and how it affects the standard deviation.

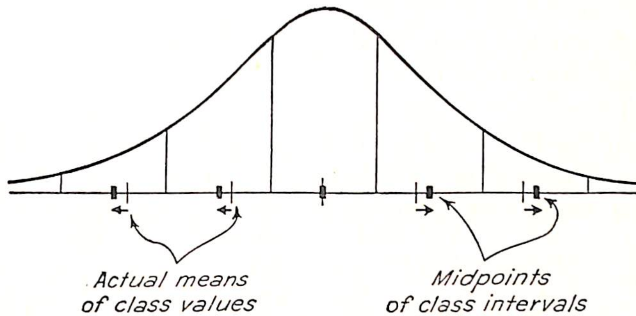


FIG. 5.5.—Illustration of grouping errors resulting from letting the midpoint of each class interval represent all cases within the interval rather than using the mean of the values for that interval. The smaller the number of intervals, the greater the error.

This phenomenon is illustrated in Fig. 5.5. There, a distribution is drawn with only five intervals. Our computations with grouped data thus far have assumed that all the values within an interval may be given a class value corresponding to the midpoint of the interval. In coarse grouping the midpoint value is not a very exact representative one because the cases are not distributed evenly, or even symmetrically, within the interval. The only exception to this is the interval that may happen to straddle the mean, in which case the midpoint and the average of the cases in the class will coincide.

In other intervals, note that the frequencies are greater toward the limit on the side nearer the middle of the distribution. If we computed an actual mean of the cases within each interval, we should find it nearer the mean of the entire sample than the midpoint is. The difference between the class mean and the midpoint of an interval is the grouping error in that interval. Above the sample mean the grouping errors are ordinarily positive (midpoint greater than the class mean) and below the sample mean the errors are ordinarily negative (midpoint less than

the class mean). The effect of the grouping errors upon the computation of a mean is usually almost nil because they are fairly well balanced. But their effect upon the average deviation, and especially upon the standard deviation, is often large enough to be concerned about. Grouping errors tend to enlarge the standard deviation and the coarser the grouping the greater is this systematic error in σ.

Sheppard's Correction.—When a correction in σ is necessary, Sheppard's formula, developed for this purpose, serves very well. When applied to a known standard deviation, it reads

$$_c\sigma = \sqrt{\sigma^2 - \frac{i^2}{12}} \qquad \text{(Sheppard's correction in } \sigma \text{ for coarse grouping)} \qquad (5.14)$$

where $_c\sigma$ = standard deviation corrected for errors of grouping.
σ = uncorrected standard deviation computed from data grouped in class intervals.
i = size of the class interval.

It is more convenient to make the correction before proceeding as far as the final step in computing σ. When σ', or when σ'^2, is known (as formula 5.10 is being utilized) we can go directly to the corrected σ by the equation

$$_c\sigma = i\sqrt{\sigma'^2 - .0833} \qquad \text{(Sheppard's correction applied to } \sigma'\text{)} \qquad (5.15)$$

To start the correction farther back in the operations, as in connection with formula (5.13), we have

$$_c\sigma = i\sqrt{\frac{\Sigma fx'^2}{N} - \left(\frac{\Sigma fx'}{N}\right)^2 - .0833} \qquad \text{(Solution of } \sigma \text{ with Sheppard's correction included)} \qquad (5.16)$$

It has been stated that when the size of class interval, i, is equal to $.49\sigma$, Sheppard's correction amounts to only about one per cent. Such an error could be tolerated unless very precise calculations are going to be done with σ after it is computed. If an interval is about one-half σ (*i.e.*, $.49\sigma$), as just stated, and if the sample is large, with a range of about 6 standard deviations, we would then have 12 class intervals. For large samples, then, 12 class intervals is a minimum for accurate computation of the standard deviation. If there are less than 12, for accurate work we should apply Sheppard's correction. Whether or not we apply this correction, therefore, depends upon the size of sample, the number of intervals, and the use we intend to make of σ.

The Standard Deviation of Combined Distributions.—There are times when we have two sample distributions, presumably from the same popu-

lation, or obtained under the same set of conditions, and we want to combine them into a single distribution. We have already seen how the mean of the combined distribution can be computed from the means of the component distributions (Table 4.7). We can also compute the standard deviation of the combined distribution from a knowledge of their standard deviations, but in doing so we also need to use their means.

TABLE 5.11.—TWO SAMPLE DISTRIBUTIONS (*A* AND *B*) AND A COMBINATION OF THE TWO (DISTRIBUTION *T*)

X	Frequencies		
	Distribution *A*	Distribution *B*	Distribution *T*
11	1	2	3
10	3	4	7
9	6	8	14
8	9	12	21
7	11	16	27
6	9	20	29
5	7	14	21
4	3	12	15
3	1	6	7
2		4	4
1		2	2
N	50	100	150
M	6.96	6.08	6.37
Σx^2	155.92	479.36	991.64
σ^2	3.1184	4.7936	4.4072
σ	1.77	2.19	2.10

Two distributions and their combination are represented in Table 5.11, also in Fig. 5.6. An examination of the figure, especially, will show that we cannot simply average the two standard deviations of the component distributions. A simple arithmetic mean of the two standard deviations (1.77 and 2.19) would be 1.98, whereas the standard deviation of the total distribution is 2.10. Even a weighted mean of the two standard deviations would not do.

If the two samples had the same mean, then deviations of all cases from the mean in either sample *A* or sample *B* would be identical with deviations from the mean in composite *T*. If the two means are different at all, this difference contributes to the dispersion of the total distribution.

Let the distance between M_a (the mean of the A distribution) and M_t (mean of the composite distribution) be called d_a, and the distance between

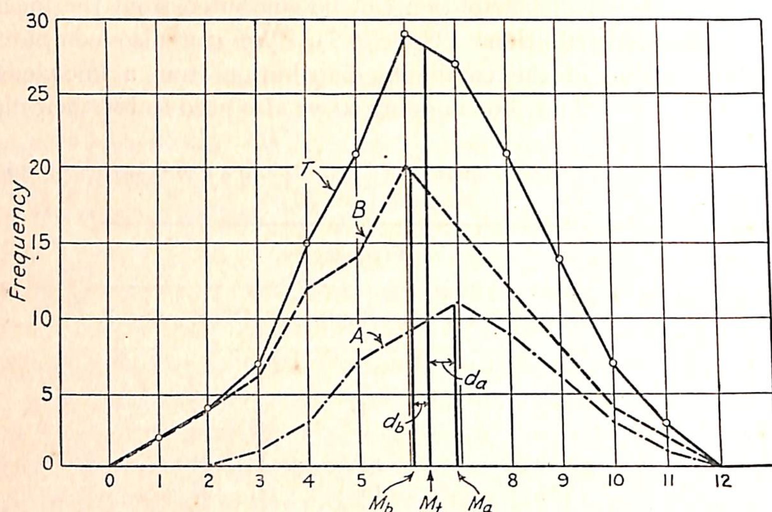


FIG. 5.6.—Two sample distributions (A and B) on the same scale of measurement and a distribution (T) of their combination. Also represented are the mean of the combined sample (M_t), of each subsample (M_a and M_b) and of the deviations of the latter from M_t (designated as d_a and d_b).

M_b and M_t, likewise, be called d_b. It has been shown that the sum of squares for the total distribution can be computed by the equation[1]

$$\Sigma x^2_t = \Sigma x^2_a + \Sigma x^2_b + n_a d^2_a + n_b d^2_b \qquad \text{(Sum of squares of combined distributions)} \qquad (5.17)$$

in which Σx^2_t = sum of squares of deviations of all measures in the total distribution from the mean of that distribution.

Σx^2_a = sum of squares of deviations of all measures in distribution A from the mean of that distribution.

Σx^2_b = similar sum for distribution B.

n_a = number of cases in sample A, and n_b the number in sample B.

d_a and d_b are as defined above, *i.e.*, $d_a = M_a - M_t$ and $d_b = M_b - M_t$.

From the equation it can be seen that if d_a and d_b equal zero (which would happen if $M_a = M_b$) then the last two terms drop out and we can say that the total sum of squares in the composite is equal merely to the summation of the sums of squares in the component distributions.

[1] For proof of this, see Appendix A.

Recalling that, in general, $\Sigma x^2 = N\sigma^2$ (see formula 5.8), we may write equation (5.17) as

$$N\sigma^2_t = n_a\sigma^2_a + n_b\sigma^2_b + n_a d^2_a + n_b d^2_b \tag{5.18}$$

Dividing both sides of this equation by N and collecting the terms,

$$\sigma^2_t = \frac{1}{N}[n_a(\sigma^2_a + d^2_a) + n_b(\sigma^2_b + d^2_b)] \quad \text{(Variance of combined distributions)} \tag{5.19}$$

Thus, we might say that the variance in the total sample is equal to a weighted average of the variances *within the component distributions* plus the weighted variances *of the sample means around the composite mean.* This formula can be extended to include any number of component samples. For each additional sample brought into the combination, there would be another expression like $n_a(\sigma^2_a + d^2_a)$. Taking the square root of both sides of equation (5.19) and extending it to include any number of components, we have

$$\sigma_t = \frac{1}{\sqrt{N}}\sqrt{n_a(\sigma^2_a + d^2_a) + n_b(\sigma^2_b + d^2_b) + \cdots + n_k(\sigma^2_k + d^2_k)}$$

(Standard deviation of combined distributions) (5.20)

in which $n_k(\sigma^2_k + d^2_k)$ stands for the last of k samples.

If the samples are all of the same size, *i.e.*, if $n_a = n_b = \cdots = n_k$, the equation reduces to the form

$$\sigma^2_t = \frac{1}{k}[(\sigma^2_a + d^2_a) + (\sigma^2_b + d^2_b) + \cdots + (\sigma^2_k + d^2_k)]$$

(Standard deviation in combined distributions of equal size) (5.21)

where k = the number of samples.

To return to the example of Table 5.11, $\sigma_a = 1.77$, $\sigma_b = 2.19$, as computed to two decimal places. For the sake of illustration, σ_t was also computed directly from the total distribution and it was found to be 2.10. Let us see whether we can arrive at the same value by the use of formula (5.20). The work is outlined in Table 5.12. There we have used the known means (to two decimal places; at least two places are necessary in most cases to give d values to more than one significant digit) to find the d values, and the two standard deviations, as given. The four contributions to the total sum of squares are given in the last column of Table 5.12. The total variance is 4.414, as compared with 4.407 when computed directly, and the standard deviation checks to two decimal places with that computed directly.

TABLE 5.12.—WORKTABLE FOR THE COMPUTATION OF THE STANDARD DEVIATION OF COMBINED DISTRIBUTIONS

Distribution	n	σ	σ^2	$n\sigma^2$
A	50	1.77	3.1329	156.6450
B	100	2.19	4.7961	479.6100
		d	d^2	nd^2
A	50	+0.59	.3481	17.4050
B	100	−0.29	.0841	8.4100
			$\Sigma x^2_t =$	662.0700
			$\sigma^2 =$	4.414
			$\sigma =$	2.10

It is probably clear that the work represented in Table 5.12 is less than that involved in setting up a total distribution and computing a standard deviation from it. It should be stated by way of warning that distributions should not ordinarily be combined if either means or standard deviations differ too much. The answer to the question, "How much is 'too much'?" cannot be given at this stage of the presentation of statistics, for the answer must depend upon standard errors of means and of standard deviations and upon tests of significance of differences, topics which appear in Ch. 9.

Standard Deviation for Augmented Values.—It should be helpful to the beginner in statistics to consider what happens to a standard deviation when we make certain systematic changes in measurements. We will consider two different changes: (1) adding a constant to each score in the sample and (2) multiplying each score by a constant. In both instances we can predict exactly what will happen. An illustration will be given to show what happens. Mathematical proofs are given in Appendix A.

Let us use a simple problem of five scores, as in Table 5.13. The same effects could be shown with any five scores we were to choose. The five scores are 11, 9, 8, 7, and 5. Their mean is 8.0 and their standard deviation is 2.0. First, let us add the constant 5 to every score. The scores so augmented are listed in column (4) and are denoted as X'. The mean of the X' scores is 13.0, which is 5 units higher than the mean of the X scores. This illustrates the fact that *if we add a constant to all scores in a sample the mean is increased by the same constant.* In terms of an equation,

$$M_{(X+C)} = M_x + C \quad \text{(Mean of } X \text{ values each plus a constant } C) \quad (5.22)$$

Notice the deviations x' [column (5)]. They are identical with the deviations x [column (2)]. Augmenting each value by adding a con-

TABLE 5.13.—THE STANDARD DEVIATION WHEN SCORES ARE AUGMENTED BY ADDING A CONSTANT OR BY MULTIPLYING BY A CONSTANT

(1) X	(2) x	(3) x^2	(4) X' $(X + 5)$	(5) x'	(6) x'^2	(7) X'' $(3X)$	(8) x''	(9) x''^2
11	+3	9	16	+3	9	33	+9	81
9	+1	1	14	+1	1	27	+3	9
8	0	0	13	0	0	24	0	0
7	−1	1	12	−1	1	21	−3	9
5	−3	9	10	−3	9	15	−9	81
Sums 40		20	65		20	120		180
Means 8.0		4.0	13.0		4.0	24.0		36.0
Standard deviations. .		2.0			2.0			6.0

stant has not changed the deviations at all. Nor has the sum of squares changed, nor the variance, nor the standard deviation, which is still 2.0. We could have changed every X by *deducting* a constant from it (which is equivalent to augmenting it with a negative amount) and the same result would follow. *When each value in a sample is increased by a constant increment, there is no change in the standard deviation.* In terms of an equation,

$$\sigma_{(X+C)} = \sigma_x \quad \text{(Standard deviation of } X \text{ values each plus a constant } C) \tag{5.23}$$

Let us next multiply each X by 3. The results are given in column (7) under the heading X''. The mean of the X'' values is 24.0, just three times the mean of the X values. The general principle is that *when all measurements are multiplied by a constant, the mean is also multiplied by the same constant.* In terms of an equation,

$$M_{CX} = CM_x \quad \text{(Mean of } X \text{ values each multiplied by a constant } C) \tag{5.24}$$

What happens to the deviations from the mean under this circumstance? The deviations x'' [column (8)] are 3 times the corresponding x deviations. The x''^2 deviations are 9 times the corresponding x^2 values. The sum of squares is also 9 times Σx^2. The variance is 36, which is 9 times the variance in the original distribution. The standard deviation is 6.0, which is 3 times that for the original distribution. The general principle is that *when all measurements in a sample are multiplied by a constant the standard deviation is also multiplied by the same constant.* This is also true when the constant is a fraction, *i.e.*, some value less than 1. The equation that describes this is

$$\sigma_{CX} = C\sigma_x \quad \text{(Standard deviation of } X \text{ values each multiplied by a constant } C) \tag{5.25}$$

These principles kept in mind will make more meaningful many things that follow later. They also explain what happens in the "short" methods of computing a mean and a standard deviation. In choosing a guessed mean, we virtually deduct the value of that guessed mean from every measurement in the sample. This operation has an effect upon the computed mean but not upon the computed standard deviation. In assigning class values (x') to the measurements, we virtually divide by a constant (i), which is equivalent to multiplying by $1/i$. The various corrections we make in applying formulas (4.4) and (5.10) bring the mean and the standard deviation back to the right levels consistent with the original measurements.

Descriptive Use of Statistics

Thus far, the chief uses proposed for measures of central tendency and of dispersion have been as simple values descriptive of total distributions. This is best appreciated when we compare different samples. As an illustration of this, see Table 5.14, in which we have a few samples of Army General Classification Test data, each based upon a different civilian occupational group. We will not concern ourselves at the moment with the question of how adequate these particular samples are either for size or for representativeness of the populations from which they are purported to come. These considerations are, of course, important if we want to generalize our conclusions to those populations. We can still compare samples as such.

Some general conclusions can be drawn from the inspection of Table 5.14. When the means and medians are placed in rank order, it will be

Table 5.14.—Statistics Describing Distributions of Scores for Selected Occupational Groups Who Took the Army General Classification Test during World War II*

Occupation	N	M	Mdn	σ	Range
Accountant	172	128.1	128.1	11.7	94–157
Lawyer	94	127.6	126.8	10.9	96–157
Reporter	45	124.5	125.7	11.7	100–157
Sales clerk	492	109.2	110.4	16.3	42–149
Plumber	128	102.7	104.8	16.0	56–139
Truck driver	817	96.2	97.8	19.7	16–149
Farm hand	817	91.4	94.0	20.7	24–141
Teamster	77	87.7	89.0	19.6	45–145

* From Harrell, T. W., and Harrell, M. S., Army General Classification Test scores for civilian occupations. *Educ. & Psychol. Meas.*, 1945, **5**, 229–240. By permission of the publisher.

seen that the occupational groups fall into an approximate rank order for socioeconomic level. It is also apparent, as should have been expected, that occupations requiring more "headwork" are highest in the list. The test emphasized verbal, reasoning, and numerical facilities.

The importance of having both means and medians lies in the information they give concerning skewness. For the lower occupational groups, particularly, the medians are slightly higher than the means. This indicates slight negative skewing. This is a somewhat surprising result, for one would expect that the higher the mean the greater the negative skewing, and the lower the mean the greater the positive skewing. When a test of moderate difficulty is administered to a group of low average ability, scores tend to bunch at the lower end of the scale (positive skewing). When the same test is given to a group of high average ability, the bunching is expected near the upper end of the scale (negative skewing). Since in the data of Table 5.14 the skewing seems to be negative for most occupational groups and most marked for those of low average ability, some explanation is demanded. We can only speculate, which means we can suggest several hypotheses which would need further investigation in order to evaluate their worth. One hypothesis might be that in any occupational group, particularly among those of lower ability in the test, a minority of the examinees were very poorly motivated or took the test under adverse conditions so that they did not perform up to their characteristic level.

Two indices of dispersion are given; the standard deviation and the total range. Each tells its own story. Standard deviations are more meaningful here if it is remembered that for the *total* range of scores, all occupational groups combined, the standard deviation was approximately 20.0. The scaling which was utilized aimed at a standard deviation of 20.0 and a mean of 100. The mean in some forms of the test turned out to be somewhat above 100. We would expect dispersions within selected occupational groups to be smaller than the dispersions for all occupations combined. With three exceptions in Table 5.14, this is true. On the whole, the higher the occupational group and the higher the mean, the smaller the dispersion. The higher groups should not be expected to scatter so far from the mean, because the mean score approaches the highest scores made by individuals in *any* group. We might expect a similar curtailment for groups with lowest means. But a study of the ranges will show why this did not occur.

The ranges, as such, are surprisingly large for all groups. It is hard to imagine any individuals in the professional groups with scores below the general average, unless those scores were low because of poor motivation

or because of advancing age, which is associated with slower rate of work. The test was a speed test. The lowest scores for the lower occupational groups are in line with expectations, but the maximum scores in those same groups are illuminating. Many a clerk or truck driver could evidently have successfully undertaken training for one of the professional occupations. In their prewar assignments they for some reason did not take full vocational advantage of their abilities. It is this fact and also the fact that men of very low academic abilities can engage successfully in the occupations like farm hand and teamster that are largely responsible for the unusually wide dispersions of scores in such occupational groups.

In this discussion we are not particularly interested in settling points concerning the relation of mental abilities to occupational level or success. The data were presented here merely as an illustration of the kind of inferences one may draw from a set of statistics and the hypotheses that may be set up for further investigation, possibly of a very fruitful nature. Such inferences and hypotheses would be impossible to make without this kind of inspection and the inspection is made possible by having the statistical information.

Uses and Interrelationships of Different Measures of Dispersion

Choice of the Statistic to Use.—Several considerations come into the picture when we decide what measure of variability to employ in any situation. One is the reliability of the statistic; its relative constancy in repeated samples. In this respect, the statistics come in the order, from most reliable to least reliable: standard deviation, average deviation, semi-interquartile range, and total range. So far as quickness and ease of computation are concerned, the four are almost in reverse order to that just given. If further statistical computation is to be given the data, such as estimating reliability of the mean and of differences between means, computing coefficients of correlation, regression equations, and the like, then the standard deviation is by all odds the one to employ.

As between standard deviation and average deviation, there is sometimes a choice. The standard deviation, because it derives from squared deviations, gives relatively more weight to extreme deviations from the mean. If a distribution should have an unusual number of extreme cases in one or both directions from the mean, some investigators prefer the average deviation to the standard deviation. This rule includes cases of markedly skewed distributions.

The semi-interquartile range gives even less importance to extreme deviations than does the average deviation and would sometimes be

given preference to both standard and average deviations for this reason. It gives more importance to the central mass of cases. When the median is the measure of central tendency adopted, Q should naturally be the companion measure of variability. Both are based upon the same principles. When distributions are truncated, or have some indeterminate values, only Q can justifiably be used to indicate variability.

To recapitulate,

1. Use the range when
   a. The quickest possible index of dispersion is wanted.
   b. Information is wanted concerning extreme scores.
2. Use the semi-interquartile range, Q, when
   a. The median is the only statistic of central tendency reported.
   b. The distribution is truncated or incomplete at either end.
   c. There are a few very extreme scores or there is an extreme skewing.
   d. We want to know the actual score limits of the middle 50 per cent of the cases.
3. Use the average deviation when
   a. There are extreme deviations which, when squared, would bias estimation of the standard deviation.
   b. A fairly reliable index of dispersion is wanted without the extra labor of computing a standard deviation.
   c. The distribution is nearly normal and we can therefore estimate σ from the AD (see formula 5.28).
4. Use the standard deviation when
   a. Greatest dependability of the value is wanted.
   b. Further computations that depend upon it are likely to be needed.
   c. Interpretations related to the normal distribution curve are desired. It will be found in a later chapter that the standard deviation has a number of useful relationships to the normal curve and to other statistical ideas.

Relationships among the Measures of Dispersion.—Previously, the standard deviation was related roughly to the range of measurements in a sample. In the general run of samples one meets in statistical work, the range varies from 4 to 6 times the standard deviation (see Table 5.8), depending upon the size of sample. If the distribution with which we deal is normal, or nearly normal, in form, we can use a number of other relationships. In a strictly normal distribution the following relationships hold:

$$Q = .845AD = .6745\sigma \qquad (5.26)$$
$$AD = 1.183Q = .798\sigma \qquad (5.27)$$
$$\sigma = 1.483Q = 1.253AD \qquad (5.28)$$

(Conversion of one measure of dispersion into another, assuming a normal distribution)

These equations are most useful for checking purposes when for some reason we have computed two or more of the statistics. They are also useful in estimating one measure of dispersion from another when we do not take the trouble to compute more than one. This should be done only with great caution, however, being assured both that the distribution is close to normal and that the one computed statistic is correct.

The Coefficient of Variation

Absolute versus Relative Variability.—Measures of variability are not directly comparable unless they are based upon the same scale of measurement with the same unit. It is even questionable whether one should compare absolute variabilities on the same measuring scale when two groups have decidedly different means. For example, the variability in height of infants might naturally be expected to be less than the variability in height of adults. If we are interested in comparing the variability in height of infants, *as infants*, with variability in height of adults, *as adults*, we need to consider infant and adult norms. These norms are naturally given in terms of means or medians. We are here concerned with *relative* variability rather than *absolute* variability. The question is more correctly stated by saying, "Is the variability of infants' heights in ratio to their mean as great as the variability of adults' heights in ratio to their mean?" We therefore need to know the ratio of the standard deviation to the corresponding mean. It is customary to multiply this ratio by 100, which tells us what percentage of the mean the standard deviation is. The formula is

$$CV = \frac{100\sigma}{M} \qquad \text{(Coefficient of variation)} \qquad (5.29)$$

Relative Variability and Weber's Law.—One important application of the coefficient of variation is in the field of psychophysics. If we ask an observer to duplicate a 90-mm. line by free-hand drawing 50 times and if we then compute the mean and standard deviation of his reproductions, we may expect a mean something like 107 mm. and a standard deviation of about 5 mm. His coefficient of variation is 4.7; or, in other words, his variability is 4.7 per cent of his mean. In duplicating a line of 180 mm. 50 times, let us say that his mean is 195 mm. and his standard deviation is 8 mm. The variability has increased as well as his average. According to Weber's law, it should have kept in step with his increase in average

and the coefficient of variation should consequently be the same. *CV* is now 4.1 per cent, or almost the same as before, but is perhaps lower than Weber's law requires. Results in the past have typically shown that with increasing mean, the *absolute* variability does increase though not so rapidly in proportion, so that the *relative* variability decreases and does not remain constant, as according to Weber's law. We are not concerned here particularly with the validity of Weber's law except as it illustrates the importance of relative variability.

When Not to Apply the Coefficient of Variation.—One important word of caution is necessary concerning the application of *CV*. *It should not be applied unless we are rather certain that our measuring scale is one of equal units and, above all, unless the absolute zero point is taken into account.* These qualifications almost entirely confine us to measuring scales with physical units, such as linear distances, weights, and time. They rule out ordinary test and examination scores, even mental-age and *IQ* units, and thus materially reduce the areas of application of *CV* in psychological investigations.

To illustrate the seriousness of this, let us note a fictitious but not unreasonable example. In a certain psychological test composed of items the mean is 8.5 and the standard deviation is 3.4. The coefficient of variation would be 340/8.5 = 40.0. The standard deviation is 40 per cent of the mean. But remember that scores on such tests do not represent distances from a meaningful or absolute zero point. Let us assume that an obtained score of zero on this test actually represents an ability that is 12 units above the genuine zero point; 12 units of the same order of magnitude of the units within the obtained range of scores. On such an "absolute" scale, the mean of the scores would be 20.5 rather than 8.5. The standard deviation would remain the same, 3.4, since we have in effect merely added 12 points to each person's score and have not disturbed the scores' relative positions. The *CV* now becomes 340/20.5 = 16.6, or less than half what it was before, while the absolute variability has remained the same.

The coefficient of variability has entered into the controversy concerning the relation of variability to learning. This has been an important problem because it was maintained that if variability increases as a group of individuals indulges in a like amount of practice in a skill or habit, there is support for the hypothesis of hereditary determination of the abilities underlying the skill. If variability decreases, it is contended that the result favors the environmentalist's hypothesis concerning determination of the abilities. Questions have arisen as to whether measures of variability should be absolute (standard deviation) or relative (coeffi-

cient of variation). Incidentally, there has been an issue as to whether time-limit or work-limit scores should be used. Results sometimes differ drastically depending upon which are used. The problems are too involved to go into adequately here. For an excellent, but brief, discussion, see Woodworth.[1]

Exercises

1. Compute the interquartile and semi-interquartile ranges for the distributions in Data 4*A*, 4*B*, and 4*F*. Interpret your findings.
2. Compute the standard deviation for any or all of the distributions in Data 4*A* to 4*F* inclusive. Use any of the formulas that seem most convenient. Report sums of squares and variances as well as standard deviations. Interpret your findings.
3. Compute the variance and the standard deviation in any or all of the distributions in Data 4*G*. Use any of the formulas that seem most convenient.
4. Compute the average deviation for any or all of the distributions in Data 4*G*.
5. Decide which measure of variability is wisest to employ with each of the distributions in Data 4*A* to 4*F* inclusive and which is second best. Give reasons.
6. In which of the same distributions would one be justified in computing a coefficient of variation and in which ones not? Give reasons.
7. Compute the standard deviation for data 5*A*, with and without Sheppard's correction. Use the various forms of the correction formula.

DATA 5*A*.—SCORES IN A FINAL EXAMINATION

Scores	*Frequencies*
70–79	1
60–69	4
50–59	10
40–49	15
30–39	8
20–29	2

8. In two distributions, *A* and *B*, the following statistics are known: $N_a = 80$, $N_b = 30$, $M_a = 42.5$, $M_b = 45.0$, $\sigma_a = 5.4$, and $\sigma_b = 4.8$ Find the sum of squares, variance, and standard deviation for a combination of these two samples.
9. Compute the coefficient of variation for each distribution in Data 5*B*. Interpret the table as it stands, and also your computed coefficients.

DATA 5*B*.—SCORES IN THREE MOTOR TESTS

Test	Tapping rate		Hand grip		Steadiness	
	Men	Women	Men	Women	Men	Women
Mean	210.4	184.0	42.1	23.9	5.64	5.13
Standard deviation	20.0	19.3	6.4	4.8	1.6	1.9
N	101	161	108	172	105	165

[1] Woodworth, R. S., Experimental psychology. New York: Holt, 1938. Pp. 173–175.

CHAPTER 6

CUMULATIVE DISTRIBUTIONS AND NORMS

Many statistical procedures, particularly those applied to test scores, are based upon the cumulative frequency distribution. Heretofore we have given frequencies as belonging to certain scores or to class intervals. In this chapter, we are interested in the number of scores or measurements falling *below* a certain point on the measuring scale. The cumulative frequency corresponding to any class interval is the number of cases within that interval *plus all those in intervals lower on the scale.*

Cumulative Frequencies and Cumulative Distribution Curves

How to Find the Cumulative Frequencies.—The cumulative frequencies are very readily found from the ordinary noncumulative frequencies. Our first example is with the already familiar ink-blot test scores (see Table 6.1). We list the scores in the first column just as before, with high scores at the top, giving in column (1) the score limits of the class intervals. We next want a single score value to assign to each interval. Where before we used the midpoint, now we choose the exact upper limit. The

Table 6.1.—Cumulative Frequency Distribution for the Ink-blot Test Data

(1) Scores in the intervals	(2) Exact upper limit of the interval	(3) f Frequencies	(4) cf Cumulative frequencies
55–59	59.5	1	50
50–54	54.5	1	49
45–49	49.5	3	48
40–44	44.5	4	45
35–39	39.5	6	41
30–34	34.5	7	35
25–29	29.5	12	28
20–24	24.5	6	16
15–19	19.5	8	10
10–14	14.5	2	2

reason is that the frequency to be given corresponding to it will be all the cases *within* the class and *below* it. All those cases fall below the exact upper limit of the class. In column (3) are given the ordinary frequencies and in column (4), the cumulative frequencies. The cumulation is started at the bottom of the list in column (3). Below the upper limit of the lowest interval (14.5) are 2 cases. Below the upper limit of the second interval (19.5) are these 2 plus the 8 in the second interval, giving 10 as the cumulative frequency. In the third interval, we find 6 cases to add onto what we already have, making 16 for the third interval. And so it goes, each cumulative frequency being the sum of the preceding one and the frequency in the class interval itself. This continues until

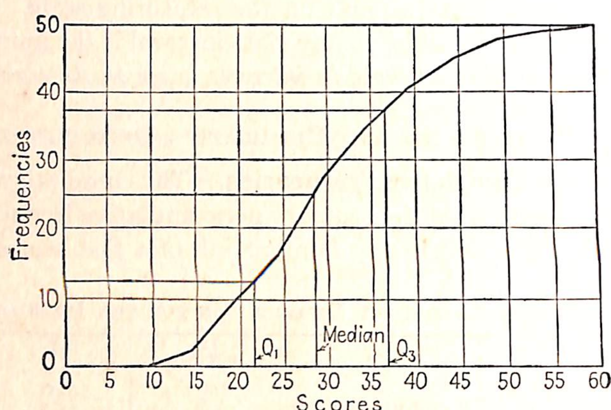


FIG. 6.1.—A cumulative frequency distribution curve for the ink-blot test.

the last (top) interval is reached. The last cumulative frequency should be equal to N (here it is 50); if not, some error has been made.

Plotting the Cumulative Distribution.—Figure 6.1 shows the cumulative frequencies we have just obtained in Table 6.1, plotted against the corresponding scores (exact upper limits). The plotting here follows much the same routine as prescribed in Ch. 3, except that here we never plot the histogram form, only the type that connects neighboring dots with straight lines. Obviously we do not obtain a polygon but rather an S-shaped curve. In order to bring the curve to the base line at the left, we assume that a zero frequency comes at the lower limit of the bottom class interval (which is the same as the top of the interval just below it). As before, the total figure is about 60 to 75 per cent as high as it is wide.

Determining Quartiles Graphically.—It is of interest to point out here the ease with which the quartiles can be graphically determined or read off the curve in Fig. 6.1. To find the median (Q_2), we first locate the

frequency of 25 ($N/2$) on the vertical axis. Draw a horizontal line over to the curve at this level. At the point where it intersects the curve, drop a perpendicular to the base line. Where this cuts the base line, read the score value. On ordinary graph paper, Q_2 can be read accurately to one decimal place. Q_1 would be similarly determined at the level of 12.5 on the frequency scale and Q_3, at the level of 37.5.

Distribution of Cumulative Percentages and Proportions.—Previously we have had reason to transform frequencies into percentages for the sake of comparing two distributions where N differs (Ch. 3). The same reason, plus more important ones, prompts us more frequently to transform cumulative frequencies into percentages. In Table 6.2, another example of

TABLE 6.2.—CUMULATIVE FREQUENCIES, PERCENTAGES, AND PROPORTIONS FOR MEMORY-TEST SCORES

(1) Scores	(2) X	(3) f	(4) cf	(5) Cumulative % cP	(6) cp
41–43	43.5	1	86	100.0	1.000
38–40	40.5	4	85	98.8	.988
35–37	37.5	5	81	94.2	.942
32–34	34.5	8	76	88.4	.884
29–31	31.5	14	68	79.1	.791
26–28	28.5	17	54	62.8	.628
23–25	25.5	9	37	43.0	.430
20–22	22.5	13	28	32.6	.326
17–19	19.5	8	15	17.4	.174
14–16	16.5	3	7	8.1	.081
11–13	13.5	4	4	4.7	.047
8–10	10.5	0	0	0.0	.000

cumulative frequencies is given. They are obtained here [column (4)] just as before. We now wish to find what percentage of 86 each cumulative frequency is. The arithmetic is simply a matter of multiplying each cumulative frequency by $100/N$. This fraction, 100/86, is equal to 1.1628. It is well here to keep a liberal number of decimal places. In Table 6.2, the cumulative percentages in column (5) are obtained by multiplying each frequency in column (4) by 1.1628. These need not be given to more than one decimal place. Sometimes it is preferable to work in terms of cumulative *proportions*, which are given in column (6).

Whereas with percentages the base is 100, with proportions the base is 1.00. Each proportion is therefore simply 1/100 of the corresponding percentage. Thus, $cp = .011628 \times cf$. The reason for using proportions will be explained later; here we shall be concerned with percentages.

The Cumulative Percentage Curve, or Ogive.—In Fig. 6.2, the cumulative percentages we have just obtained in Table 6.2 are plotted as points against the corresponding score points (exact upper limits of class intervals). Again, an S-shaped curve results. Now that it is standardized as to height, it is sometimes called an *ogive*. *The ogive is*, in other words, *the cumulative percentage distribution curve.*[1] Two ogives are much more readily compared than two ordinary cumulative curves because of their common height. But this is not the only use of an ogive, as we shall soon see.

Centile Norms

Finding Centile Points by Interpolation.—*A centile point, or centile, is a value on the scoring scale below which are any given percentage of the cases.*[2] For example, the 90th centile is the point below which are 90 per cent of the scores, and the 24th centile is the point below which are 24 per cent of the scores.[3]

Deciles and Tenths.—We have already seen how to interpolate in order to compute a median and other quartiles. Actually, the median is at the 50th centile, Q_1 is at the 25th centile, and Q_3 is at the 75th centile. It is but a step further to generalize this to any centile one desires. We could choose to interpolate any centile; the 63d, the 81st, or the 8th. Our interest in testing happens to stress the centiles that are multiples of 10—the 90th, 80th, 70th, etc., down to the 10th. These are called the *deciles*, for they divide the distribution into tenths, just as the quartiles divide it into quarters and the median, into halves.

The Process of Interpolation.—The principle of interpolating is not new. Table 6.3 shows how we may work out the deciles systematically. The complete headings of the table make the work almost self-explanatory, but let us follow through one or two examples. First we need to know how many cases out of the total of 86 we need to include in any given

[1] The ogive may also be in terms of cumulative proportions, since proportions and percentages are used interchangeably.

[2] The term *centile* is often called (superfluously) *percentile* in the literature. There is about as much excuse for speaking of *perdecile* or of *perquartile*.

[3] The term *centile,* without reference to a scale of measurement, should mean *centile rank*. Thus, to say that an individual is at the 24th centile indicates his rank among a hundred persons. Being better than 24 of a hundred, he would rank 25th from the bottom.

percentage. Ninety per cent of 86 is 77.4, which we find in column (2). We must count up the scoring scale among the frequencies until we include 77.4 cases. Reference to Table 6.2 shows that we get by accumulation 76 cases up to the score point 34.5. We need 1.4 more cases

TABLE 6.3.—CALCULATION OF CENTILES, OR CENTILE POINTS BY INTERPOLATION IN THE MEMORY-TEST DATA

(1) Percentage below the centile point	(2) Number of cases below the centile point	(3) Cumulative frequency actually below the interval containing the centile point	(4) Lower limit of interval containing the centile point	(5) Distance of centile point above lower limit	(6) The centile point
90	77.4	76	34.5	+ $\frac{1.4 \times 3}{5}$	35.3
80	68.8	68	31.5	+ $\frac{.8 \times 3}{8}$	31.8
70	60.2	54	28.5	+ $\frac{6.2 \times 3}{14}$	29.8
60	51.6	37	25.5	+ $\frac{14.6 \times 3}{17}$	28.1
50	43.0	37	25.5	+ $\frac{6 \times 3}{17}$	26.6
40	34.4	28	22.5	+ $\frac{6.4 \times 3}{9}$	24.6
30	25.8	15	19.5	+ $\frac{10.8 \times 3}{13}$	22.0
20	17.2	15	19.5	+ $\frac{2.2 \times 3}{13}$	20.0
10	8.6	7	16.5	+ $\frac{1.6 \times 3}{8}$	17.1

among the 5 in the next higher interval. There are 3 score units in the interval; so we have to proceed 1.4/5 times 3, or, as given in columns (4) and (5) of Table 6.3, we add to 34.5 the amount $\frac{1.4 \times 3}{5}$, which gives 35.3 as the centile point. We say that P_{90} (90th centile) equals 35.3. To take a second example, let us solve for P_{10}. Ten per cent of 86 is 8.6. Counting up to a score point of 16.5, we find 7 cases, which leaves us needing 1.6 more out of the 8 in the next interval. P_{10} is therefore equal

to $16.5 + \frac{1.6 \times 3}{8}$, which equals 17.1. The remaining centiles are similarly determined and are listed in the last column of Table 6.3.

The Utility of Centile Norms.—Test scores of various kinds are frequently interpreted in terms of centile norms, for very good reasons. In the first place, a raw score of so many points means very little to us. Tell a student's adviser that his advisee made a score of 59 points in an algebra-achievement examination, 175 points in an English-achievement examination, and 121 points in a general scholastic-aptitude test, and without further information the adviser does not know whether his advisee is low in all tests, high in all tests, or low in one or two and high in the remaining. But tell him that a score of 59 points in algebra is at the 99th centile, the 175 points in English is at the 32d centile, and the 121 in scholastic aptitude is at the 48th centile, when those centiles were established by the scores from 1,500 freshmen entering the University with the advisee in question; then he will have some usable information. The student in question is extremely high in algebra, moderately low in English, and about average in general scholastic aptitude. The chief utility of centile norms is (1) to give some conception of the general level of a score in a known population, and (2) to put scores from different tests on a comparable basis.

Finding Centile Norms by Interpolation.—If we wished to have a table of centile norms for the memory test, we could now use the nine decile points already found by interpolation as they are listed in the last column of Table 6.3. Then when a student came along with a score of 22 we could say that he is at the 30th centile; another student with a score of 30 is at the 70th centile, etc. When a score came up that is not exactly listed we could find its centile equivalent by interpolation. For example, a score of 21 would be at the 25th centile, and a score of 27 would be at about the 53d centile.

Centile Norms from Smoothed Ogives.—But there are objections to the use of interpolated centiles as norms. Chance irregularities in distribution from a small sample often give a distorted picture of the true situation that probably obtains in the larger population. After all, it *is* the larger population that we wish to represent in our norms, or at least we should like to compare future individuals' scores with something more stable and general than our limited sample. For this reason the author strongly recommends that centile norms be set up in terms of the smoothed ogive. Interpolated norms are derived from the unsmoothed curve and, as was said, they are affected by minor irregularities that are probably a peculiarity of this sample only and not of the general population. The

smoothed ogive may be taken as an estimation of the distribution of the general population of which our group is a sample. When a sample is large, very little smoothing is necessary. Even with small samples, at times surprisingly little smoothing need be done.

In Fig. 6.2, a smoothed ogive (by inspection and free-hand drawing) has been drawn. The aim is to bring it as close as possible to all points, and if points must be untouched by the curve, there should be about as many below the curve as above it. If too glaring discrepancies occur between points and curve after smoothing, it is probably best to discard the attempt to use these data as a basis for norms or else to add more cases until sampling irregularities are greatly reduced.

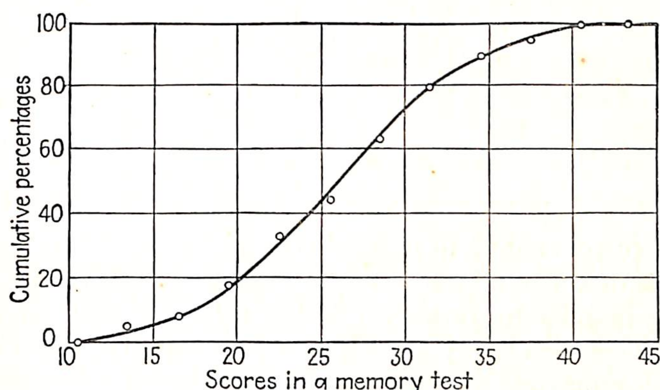


FIG. 6.2.—Smoothed cumulative distribution curve for the memory-test scores. Frequencies are in terms of percentages.

Reading Centile Scores from a Graph.—Having satisfied oneself as to the smoothed ogive, the next step is to read off the diagram the score points corresponding to the centile ranks for which norms are required. For this purpose the diagram should be enlarged sufficiently for easy reading and the graph paper finely ruled so that score points may be accurately read to one decimal place. In Table 6.4 are given the score points corresponding to centiles 10 to 90, as before, but also to 95 and 99 at the upper end and to 5 and 1 at the lower end. The reason for including these extra points at the extremes is that there is actually a great range of ability above the 90th centile and also below the 10th centile. In fact, the range of ability is about as great beyond the 90th centile as it is between the mean and the 90th centile, and as great below the 10th centile as between that point and the mean, when the distribution is normal.

A Defect in Decile Scales.—One defect of the centile scale, as a measuring scale, is that it exaggerates individual differences, relatively, near the center of the distribution as compared with those near the ends. Giving

TABLE 6.4.—CENTILE NORMS FOR THE MEMORY TEST, DERIVED FROM THE SMOOTHED OGIVE

Centile	Score point	Integral score
99	40.5	41
95	37.1	38
90	34.9	35
80	31.8	32
70	29.5	30
60	27.9	28
50	26.1	27
40	24.3	25
30	22.5	23
20	20.4	21
10	17.5	18
5	14.9	15
1	11.9	12

score norms corresponding to selected centiles beyond 10 and 90 compensates for this defect to a large extent. Because of this same defect, it is not the best practice to work with decile norms, for to do so often leads the user of the norms to lay too much stress upon differences among the great average group and too little upon those where tests discriminate best.

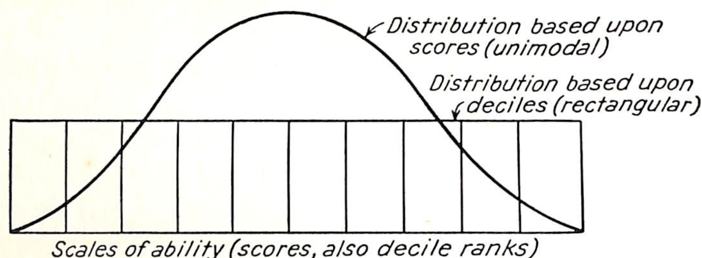


FIG. 6.3.—Showing how, when a distribution is converted to one with decile units on the base line, a distribution that was unimodal, and perhaps normal, becomes rectangular. The areas of the two distributions are approximately equal.

Figure 6.3 illustrates how a decile scale distorts differences along the scale. This figure is so drawn that the 10 decile divisions cover the same total range as the original scores. The heights of the rectangles are drawn so that the total area in the 10 categories combined is equal to that under the original curve. The new frequency distribution, when decile ranks are given equal distances on the measurement scale, is rectangular. It is as if we had pressed down upon the center of the original distribution,

forcing the central individuals farther apart, and to make up for it, we group individuals who are spread over the tails of the original curve into narrower categories.

Another illustration of the distorting effect of decile and centile scales when we give equal distances to numerically equal intervals is shown in Fig. 6.6. Here are shown parallel scales for the memory test. Corresponding centile ranks and raw scores are connected by dotted lines. From this it will be seen, in another way, how raw-score distances near the center become relatively spread and how equal distances near the extremes are relatively condensed when converted to centile-rank values.

It is probably best that decile norms, as such, be consigned to the limbo of forgotten procedures. In their place the author recommends the use of a *C* scale, which will be described in a later chapter (Ch. 12). Centile norms will continue to be useful, but it is urged that they be constructed in a way that will give more correct impressions of scale positions, as will now be described.

Integral Centile Points.—Before doing that, however, a further word of explanation of Table 6.4 is in order. The last column of "integral scores" is merely a revision of the second column by way of rounding to whole numbers. Tables of norms are frequently given in terms of whole numbers, mainly because scores are obtained as whole numbers. We should say that an obtained score of 41 is better than 99 per cent of the group can make, and a score of 18 is better than only the lowest 10 per cent can make. It should be noticed that *every fractional score is rounded upward to the next whole number;* thus 37.1 becomes 38. Since an obtained score of 37 covers a range of 36.5 to 37.5, more than half of those making this score would *not* be better than 95 per cent. The first score, counting from below upward, that is *totally* better than 95 per cent is a score of 38. This is why, in this and in other cases in this table, we round upward to the next higher integer.

A Graphic Profile Chart.—Many profile charts based upon centiles show graphically the deciles at equidistant levels along the scale. This gives an erroneous conception of the relative spacing of ability or talent, as was pointed out in a preceding paragraph. Actual differences in ability are probably more accurately indicated by the raw-score units than they are by centile-rank units, which relatively magnify the central portions of the distribution. If it is assumed that the actual distribution for the norm group is Gaussian, or normal, in shape, the relative spacing of the various centiles that we customarily include in our norms should be as given in Table 6.5. In the first column are the customary centile ranks. In the second column are the corresponding distances from the mean (and

TABLE 6.5.—THE DISTANCE OF CENTILES FROM THE MEAN IN NUMBER OF STANDARD DEVIATIONS IN A NORMAL DISTRIBUTION

Centile, rank	*Number of Sigmas from the Mean*
99	+2.33
95	+1.64
90	+1.28
80	+0.84
70	+0.52
60	+0.25
50	0.00
40	−0.25
30	−0.52
20	+0.84
10	−1.28
5	−1.64
1	−2.33

median) when the standard deviation of the distribution is adopted for convenience as the unit. The corresponding centile ranks and sigma distances are also represented in Fig. 6.4. The correspondence of deviation from the mean with centile rank depends entirely upon the mathematical

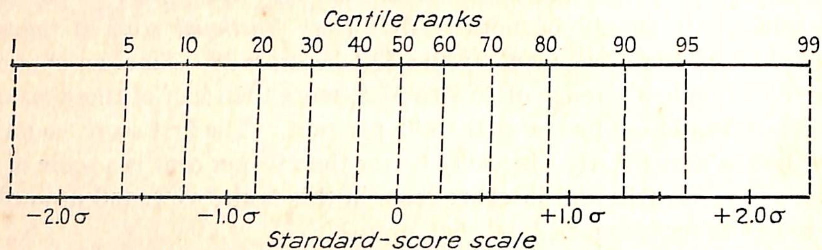


FIG. 6.4. Showing on parallel scales standard scores and corresponding centile ranks. Since standard scores are given equal spacing, centile ranks have unequal spacing. Had centile ranks been given equal spacing, standard scores would have had unequal spacing.

relations that hold true for the normal distribution curve, and the reasons for this need not concern us here. The author merely proposes to use this spacing of the centile ranks in setting up a profile chart and has done so in Fig. 6.5.

Here, in Fig. 6.5, each centile is drawn at a distance from the mean proportional to its corresponding sigma distance given in Table 6.5; *i.e.*, centiles 99 and 1 are 2.33 sigma units from the mean, centiles 90 and 10 are 1.28 units away, etc., though those distances are not labeled numeri-

cally in the chart and need not be. Once having located them at the proper distances, we may forget the sigma values.

Provision has been made for four tests in the profile chart, the memory test whose norms we have determined in previous parts of this chapter, a vocabulary test, a word-building test, and a sentence-construction test, whose norms were determined elsewhere. For the memory test, the integral scores have been written in at their corresponding centiles, being guided by the list of score points in column (2) of Table 6.4. Once the scores nearest those points are located and written in the diagram, the other, intervening scores can be introduced. The same was true for the other test norms, though because of crowding, some integral scores have been omitted. The student whose profile is shown earned raw scores of 28, 88, 20, and 23, respectively, in the four tests. Those four scores have been encircled and then connected with straight lines to complete the profile. We can now see the general trend of this student's ability in these four tests taken together, and we can read off his centile rating in each test at a glance. Furthermore, a much more accurate conception of his fluctuation in ability is given than would have been true in a diagram with equidistant deciles.

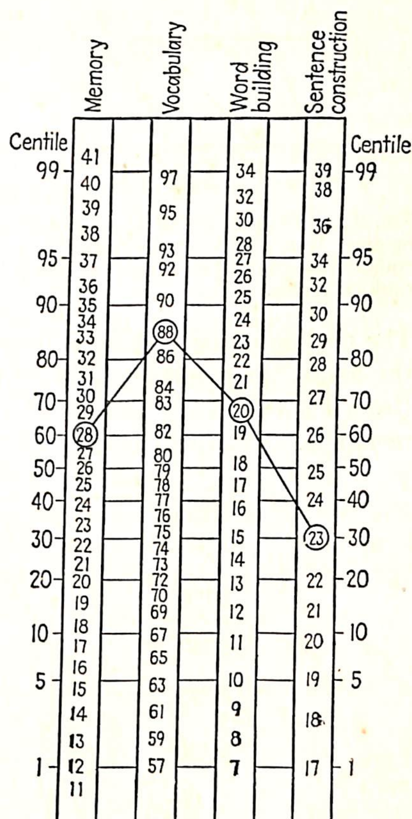


FIG. 6.5.—An example of a profile chart based upon centile norms. Note that the centile ranks are not spaced at equidistant intervals but at intervals based upon corresponding sigma distances from the mean (see Table 6.5 and Fig. 6.4).

Figure 6.6 shows how, if we had spaced the centile ranks at equidistant intervals, as is sometimes done, the corresponding separations on the score scale would have been very unequal in different parts of the scale. As a general principle, individuals are best discriminated by tests where they are spread thinnest in the distribution.

A Bar Diagram of Distributions of Scores.—A useful graphic device for picturing distributions of scores is shown in Fig. 6.7.[1] The bar diagrams

[1] Similar diagrams have been used for some time by the Cooperative Test Service.

there illustrate the distributions of three groups of students who were taught by three different instructors but who were given the same final examination, an objectively scored achievement examination in English.

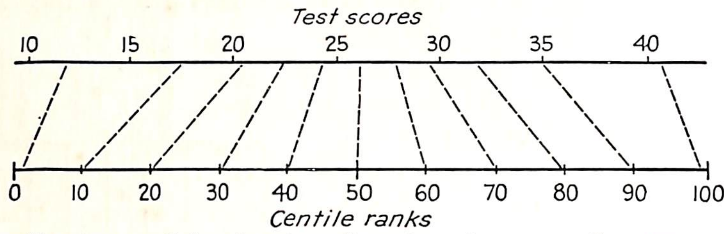


FIG. 6.6.—Showing parallel scales of centile ranks and corresponding raw scores on the memory test. Here centile ranks are equally spaced on their scale and raw scores are equally spaced on their scale. Unequal raw-score intervals correspond to equal centile-rank intervals.

The median of each group is marked by a short horizontal line through the bar at the median-score level. The range of the middle 50 per cent (from P_{25} to P_{75}, or from Q_1 to Q_3) is shown in each case by the open rectangle. The black bars extend out to the points P_{10} and P_{90}—in other

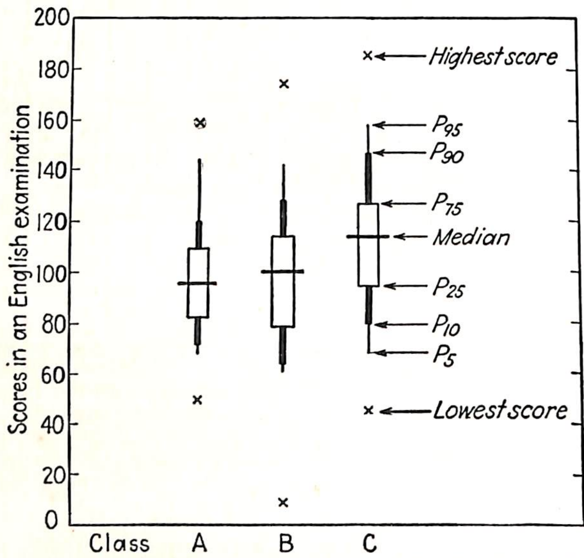


FIG. 6.7.—A graphic device for visual comparison of distributions, showing important centile values and total ranges.

words, to include the middle 80 per cent of the cases. The lines extend to points at P_5 and P_{95}, or to include the middle 90 per cent of the cases. The highest and lowest single scores are marked by the small x's. Thus several meaningful centile points are labeled, as well as the entire range.

Interpretation of Bar Diagrams.—One important use of bar diagrams is the ready comparison of groups that they afford. In Fig. 6.7, for example, it is obvious that the three medians come in the order 1, 2, 3 for groups *C*, *B*, and *A*, respectively. The variabilities of the three groups come in the order *B*, *C*, and *A* when we depend upon total ranges. The groups come in almost the same rank order for variability when we compare ranges of middle 90 per cent, but again the order *B*, *C*, *A* is probably correct in comparing middle 50 per cents, though *B* and *C* are very close together in this respect. As to topmost scores, they come in the same order as for medians *C*, *B*, *A*, but for bottom scores the order is *A*, *C*, *B*. As to skewness, the most symmetrical distribution, all things considered, is probably that for group *B*, and the least symmetrical is for group *A*, which is positively skewed. The special virtue of this kind of comparison, as contrasted with that afforded by means of frequency polygons and ogives, is that many more facts about a distribution can be recorded, and yet because of no overlapping of the drawings there is direct comparison without confusions.

Exercises

1. Carry through the following steps for the first distribution of chemistry-aptitude scores in Data 3*C* (Ch. 3).
   a. Find the cumulative frequencies, and tabulate them.
   b. Plot a cumulative distribution curve similar to Fig. 6.1.
   c. Find the cumulative percentages and proportions, and tabulate them.
   d. Plot the ogive distribution, showing the smoothed curve.
   e. Compute the interpolated centiles that divide the distribution into tenths.
   f. Derive centile norms from the smoothed ogive, and set up a table of norms.
   g. Prepare a centile profile chart including the norms for this test and for one or two others for which you have data.
2. Repeat the steps, particularly *a*, *c*, *d*, and *f*, for any other distribution of test scores.
3. Prepare bar diagrams like those in Fig. 6.7 for comparing two or more distributions, such as the two in Data 3*C*, or Data 4*F* (Ch. 4).

CHAPTER 7

THE NORMAL DISTRIBUTION CURVE

Repeatedly have sets of measurements in psychology and education yielded frequency distributions that resemble the bell-shaped normal, or Gaussian, curve. Because the normal curve has so many useful mathematical properties, it is quite natural that we should exploit those properties in dealing with psychological and educational data. Without the use of the Gaussian curve and its convenient characteristics, many things that we now do with data would otherwise be impossible. It is important, therefore, that the student develop at least a moderate understanding of the normal curve in order that he may wisely apply the statistical procedures that depend upon it.

Normality of Distribution Is Assumed.—It must be confessed at the outset that no set of data ever obtained, whether they be measurements of a group of individuals with respect to some biological, psychological, social, or educational trait or whether they be repeated observations of a single phenomenon, ever conforms exactly to the normal distribution pattern. Even though the larger population from which our sample came is perfectly normally distributed (even this is probably never strictly true), sampling, no matter how extensive or representative it may be, is bound to give us some irregularities, with deviations from the normal form. Whenever, therefore, we treat our data as if they were normally distributed, or arose from a population that is normally distributed, we are assuming an ideal pattern for the sake of simplicity, rationality, and convenience. Sometimes we are more justified and sometimes less; we can never be absolutely sure, because the entire population is rarely or never measured, and the true shape of distribution is never known.

We can justify our assumption of normality in several ways. One is the rational approach, which attempts to point out that the phenomenon we are measuring results from a number of independent causes occurring in chance combination, as in the tossing of coins or in the combinations of nonlinked hereditary genes. Very rarely is this kind of argument possible because of our ignorance of underlying causes. Another kind of approach is empirical, in which we can show that, with the use of the measuring scale that we did use, the grouped data present a frequency

distribution that obviously possesses a bell-shaped contour. Furthermore, there are statistical tests that can be applied to show whether or not the frequencies we obtained deviate so much from the normal-curve picture as to cause us to reject our hypothesis that the data came by random sampling from a normally distributed population.

Two Reasons for Caution.—There are two considerations, however, which should cause us to pause before making the hypothesis or assumption of normality. One has to do with the question of sampling and the other with the question of the correctness of our measuring scale. A population may well be normally distributed, yet because of our method of drawing cases for measurement we may obtain a skewed or otherwise distorted form of distribution. This is a case of *biased sampling*. A large population of ten-year-old children would probably be distributed normally when measured for mental age. But if we confine ourselves to ten-year-old children in the fourth grade only, where most ten-year-olds are probably present because of mental retardation and a few for other reasons, the distribution of mental ages would be positively skewed. The ten-year-olds in the sixth grade would probably yield a negatively skewed distribution, for the majority of them are accelerated by reason of precocity and a few for other causes. Both are cases of biased sampling. An unbiased, representative sampling would not confine itself to fifth-grade children, but would take ten-year-olds in correct ratios from all grades where they appear, would take them in correct proportions as to sex, economic status, and other factors considered significant.

When a test or examination is used as the measuring instrument, the form of distribution of scores will depend upon many factors other than the form of distribution of the population. One of these factors is the level of difficulty of the test relative to the level of ability of the population. Even if the population is normally distributed in the ability measured, unless the test is of an appropriate level of difficulty a normal distribution of scores in a sample will not be obtained. If the test is too difficult, the distribution will be positively skewed, like that labeled A in Fig. 7.1. If the test is of moderate difficulty for the group, a symmetrical distribution like that labeled B will occur. If the test is too easy for the group examined, the distribution will be negatively skewed, like C in Fig. 7.1. Other degrees of skewing might occur. The effect of skewing, when we are sure that the correct form of distribution should be symmetrical, may be regarded as a systematic distortion of the scale of measurement. The too difficult test tends to make the numerical units among the low scores stand for relatively large intervals of ability, and the too easy test to make the units among the high scores also stand for

relatively large intervals. This principle should be clear from a study of Fig. 7.1.

Other factors than difficulty may distort sample distributions. Later (Ch. 17) it will be shown how degree of reliability of scores may affect the form of distribution, causing tendencies toward sharpness of the rise in the center versus flatness, tendencies toward bimodality, and even U-shaped distributions. Another distorting factor may be the unsuitability of the scale. As was pointed out in an earlier chapter (Ch. 4), work-limit scores and time-limit scores tend to be reciprocals of each other. If the one kind of score in a task is normally distributed, the other will probably not be.

These cautions kept in mind should serve to inhibit dogmatic assertions that might otherwise be made about the shape of a distribution. The shape of a distribution is always a function of the kind of measuring scale,

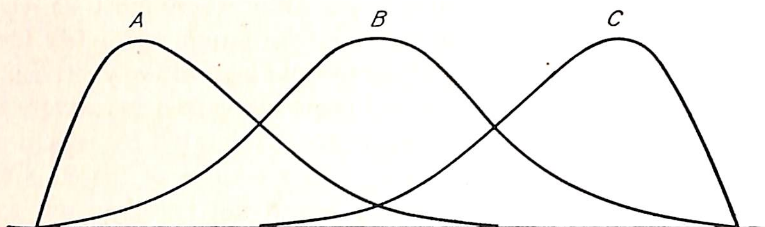


FIG. 7.1.—Showing how a test at three different levels of difficulty may yield distributions of raw scores differing markedly in skewness regardless of the form of distribution of ability in the population.

and all conclusions that involve form of distribution should take this fact into account. The conviction that general populations are genuinely normally distributed with respect to most qualities is very strong, however; so it is usually the marked deviation from normality in a sample that arouses questions. We may then question either our method of sampling or our measuring scale. One or both of these factors may be responsible for the discrepancy. But when our sample distribution turns out reasonably normal in appearance, because of the conviction just mentioned we may feel some assurance that our sampling and our measuring scale are probably free from distortions, though of course we can never be certain of this. The conviction does lead us to apply the Gaussian curve in many useful ways, even in turning crude judgments into scaled measurements, as we shall see later (Ch. 19). We frequently feel that the risk in making the normal assumption is well worth while because of the invaluable results and conclusions it affords. We can always state our conclusions with the reservation that they are true to the extent that our assumptions are valid. As a matter of fact, all other conclusions

should be couched in similar terms, for none is without its foundation of assumptions of one kind or another, whether stated or not. All scientific conclusions rest on assumptions, in the final analysis, and he who would know the import of those conclusions best is the one who knows those assumptions best.

The Nature of the Normal Curve

The Relation of the Normal Curve to Probability.—The Gaussian curve is also sometimes called the *normal probability* curve and is said to be the result of the "laws of chance." In a sense, this is true. We cannot here go into an involved discussion of probability and of the way in which the Gaussian curve is logically related to probability. It is sufficient for our present purposes to point out the usual example of how a normal distribution can be approximated by means of coin tossing. If we thoroughly shake a set of 6 coins and toss them to land where and how they may, the result can turn out in seven different ways; the number of heads can vary all the way from 0 to 6. In a total of 64 tossings, according to the principles of probability, we should expect the following frequencies for various numbers of heads:

Heads	0	1	2	3	4	5	6
Frequencies	1	6	15	20	15	6	1

If we tossed the 6 coins twice as many times, we should expect these frequencies to be doubled. Actually obtained frequencies will deviate from these expected ones by small amounts. In one such experiment with 128 tosses, the obtained frequencies were as given here:

Heads	0	1	2	3	4	5	6
Obtained frequencies	2	14	25	38	36	12	1
Expected frequencies	2	12	30	40	30	12	2

This situation is shown graphically in Fig. 7.2, where the obtained frequencies furnish the basis for the histogram and the expected frequencies furnish the basis for the superimposed normal curve.

A 6-coin problem gives us a 7-sided frequency polygon (not counting the base line). A 10-coin problem gives us an 11-sided contour, etc., the number of sides being equal to the number of coins plus 1. If we do not enlarge the base line of our distribution but keep subdividing it into smaller and smaller units as we increase the number of coins, the contour of the distribution curve approaches the smooth bell form. The num-

ber of class intervals we choose in grouping obtained measurements has nothing to do with the number of coins, our choice being entirely arbitrary. The class intervals and their frequencies merely give us descriptions of the contour at points along the way. If there are things like coins in the phenomenon we are measuring (*i.e.*, "coins" such as genes, which may be present or absent, or such as responses that do or do not occur) we almost always lack information as to how many such "coins" are operating. Probably there are a great many, although even if there were only 6, as in the coin example, and if our measurements naturally fell therefore

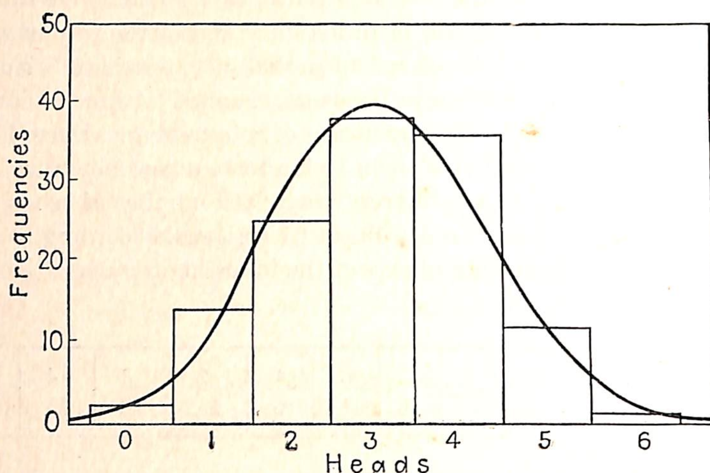


Fig. 7.2.—A distribution curve representing the frequencies with which various numbers of heads are expected by chance in tossing six coins; also, in histogram form, the obtained distribution from 128 tossings.

into seven class intervals, the normal distribution could still be roughly approached, as can be seen in Fig. 7.2.

The Equation for the Normal Curve.—Mathematically, when we are dealing with the properties of the normal curve, it is the situation with an infinite number of "coins" that we suppose. This enables the mathematician to give to the curve an equation that describes the relationship of a frequency to its corresponding measurement. This equation reads

$$Y = \frac{N}{\sigma\sqrt{2\pi}} e^{\frac{-x^2}{2\sigma^2}} \qquad \text{(Equation for the Gaussian or normal curve)} \qquad (7.1)$$

where Y = frequency.
N = number of measurements.
σ = standard deviation of the distribution.
π = 3.1416.

e = 2.718 (the base of the Naperian system of logarithms).

x = deviation of a measurement from the mean (or $X - M$).

Since the values for π and e are known, if we substitute them in the equation, it becomes

$$Y = \frac{N}{2.5066\sigma} 2.718^{\frac{-x^2}{2\sigma^2}}$$

For any distribution we may have at hand, we know the values for N and for σ, and these can be inserted in their places in the equation. The equation would then be in a form with only Y and x the unknowns. We could then assign certain values to x, within the range of our measurements, and then solve the equation for the corresponding values of Y. In this way, we could determine the entire normal distribution curve that best fits our data. The arithmetical work would be rather laborious. Fortunately, we have the use of statistical tables to aid us in this. Table B, in Appendix B, is one well suited to this purpose.

Determining the Best-fitting Normal Distribution for a Set of Data.—For the sake of an illustration that will help us to appreciate the meaning of the normal curve, let us find the expected frequencies in a particular instance, a distribution of 86 scores in a memory test. The best-fitting normal curve for any set of data has the same mean and standard deviation as those computed from the actual data. The distribution of obtained frequencies of memory-test scores is given in column (7) of Table 7.1. The mean of this distribution is 26.1, and the standard deviation is 6.45. Our task is to find the frequencies to be expected in the same class intervals for a normal distribution with a mean of 26.1, a standard deviation of 6.45, and an N of 86..

Standard Measurements or Scores.—In order to use equation (7.1) to find these frequencies, we must know how far each class interval deviates from the mean in terms of standard deviations. Each interval is given the value of its midpoint as its point on the score scale X. These X values are listed in column (2) of Table 7.1. Note that we have included one class interval beyond the range of obtained scores at each end of the distribution. This is because the best-fitting normal curve usually has some small frequencies (perhaps fractional) in those extreme positions, even though the obtained frequencies there are zero. The equation for the normal curve calls for *deviations* rather than original scores—in other words, for $X - M$, or small x, for each class interval. These are listed in column (3). In this problem, each one is found by the solution of $X - 26.1$ for every interval. A simple check is to see that each one is three units (the size of the interval) distant from its immediate neighbors.

TABLE 7.1.—OBTAINING THE EXPECTED FREQUENCIES f_e IN THE CLASS INTERVALS FOR THE MEMORY TEST, ON THE ASSUMPTION THAT THE TRUE DISTRIBUTION IS NORMAL

(1) Scores	(2) X Midpoint	(3) x Deviation	(4) z Standard score	(5) y From Table B	(6) f_e Expected frequency	(7) f_o Observed frequency
44–46	45	+18.9	2.93	.0055	0.2	0
41–43	42	+15.9	2.47	.0189	0.8	1
38–40	39	+12.9	2.00	.0540	2.2	4
35–37	36	+ 9.9	1.53	.1238	5.0	5
32–34	33	+ 6.9	1.07	.2251	9.0	8
29–31	30	+ 3.9	0.60	.3332	13.3	14
26–28	27	+ 0.9	0.14	.3951	15.8	17
23–25	24	− 2.1	−0.33	.3778	15.1	9
20–22	21	− 5.1	−0.79	.2920	11.7	13
17–19	18	− 8.1	−1.26	.1804	7.2	8
14–16	15	−11.1	−1.72	.0909	3.6	3
11–13	12	−14.1	−2.19	.0363	1.5	4
8–10	9	−17.1	−2.65	.0119	0.5	0
Sums....					85.9	86.0

Each column of numbers is derived from the one preceding by the following computations (see text for explanations):
Column (3): $x = X - 26.1$.
Column (4): $z = x/6.45$.
Column (5): y comes from Table B.
Column (6): $f_e = 40 \times y$.

The next step involves a new process; the determination of the *standard measurement or standard score*, for every interval. The standard score is given by the formula

$$z = \frac{x}{\sigma} = \frac{X - M}{\sigma} \qquad \text{(A standard score or measure)} \qquad (7.2)$$

In the equation for the normal curve, it will be seen that the exponent of e, which is $-x^2/2\sigma^2$, can be written $-(\frac{1}{2})(x/\sigma)^2$, or in other words, it is $\frac{1}{2}$ times the standard score squared. We shall find the standard score invaluable again and again. The statistical tables are constructed on the basis of standard scores. It matters not, then, what our original means and standard deviations are numerically. Reducing all raw scores to standard scores places them all on the same basis or common denomi-

nator. For our illustrative problem, the standard scores are given in column (4) of Table 7.1. Each number in column (4) is obtained by dividing the corresponding number in column (3) by 6.45, the standard deviation.

Determining Frequencies for the Class Intervals.—Having obtained the standard score for each class interval, we are now ready to look up the corresponding ordinate in the general statistical table, Table B. These are listed in column (5) of the work table. The ordinates in this table are not exactly the frequencies we have been wanting to find. Those frequencies also depend upon N [see equation (7.1)]. Table B is constructed on the assumption that $N = 1$, and $\sigma = 1$. For our distribution of 86 cases and a different σ, we must make a certain adjustment. We must multiply each y value by a certain number to find the expected frequency f_e. The general formula is

$$f_e = \left(\frac{iN}{\sigma}\right) y \qquad \text{(Expected frequency in a best-fitting normal distribution)} \qquad (7.3)$$

In this problem,

$$\frac{iN}{\sigma} = \frac{3 \times 86}{6.45} = \frac{258}{6.45} = 40.0$$

When this multiplier is used with the numbers in column (5), the frequencies we desired are finally forthcoming, and they are given in column (6).

Formula (7.3) may be made to appear reasonable if we look at it in the following manner. The expected frequencies (f_e) must be of the order of magnitude of the obtained frequencies (f_o). The sum of the obtained frequencies is, of course, equal to N. The expected frequencies are, therefore, proportional to N, as formula (7.3) states. They must also be proportional to the size of class interval (i) because the *larger* the size of interval, the *smaller* the *number* of them, and, since they add up to N, the larger each frequency is. The appearance of σ in the denominator is not quite so easily explained. It is best explained when we consider the equation for the normal curve. Ignoring the expression involving e (with its exponent) in equation (7.1), we find that Y is proportional to $N/\sigma\sqrt{2\pi}$. When we let both N and σ equal 1, as is the case in the tables on the normal curve, y is proportional to $1/\sqrt{2\pi}$. From this we see that the ratio of Y to y is equal to N/σ. Thus, from another approach we can account for the presence of σ in formula (7.3) as well as the presence of N.

Comparing Obtained and Theoretical Frequencies.—As a rough check upon all the work, we sum the expected frequencies, and the result should be very close to N but will usually be slightly less than N, because in

the normal curve there are still fractions of frequencies even beyond the limits we have included here. Had we not gone one class interval beyond the obtained data, we should have lost .2 of a frequency at the upper end and .5 at the lower, and the sum would have been 85.2 instead of 85.9. As it is, we have still lacking only .1 of a case; not enough to worry about, and we may accept our check as one indication of correct work. A comparison of expected with obtained frequencies is always a rough check but is very rough, because we expect small discrepancies within class intervals. Looking down the columns, we find only one or two serious discrepancies. One is the difference between 15.1 and 9, and the other is between 1.5 and 4. Both the obtained frequencies of 9 and 4 are out of line but are probably merely chance discrepancies, coming under the heading "errors of sampling," and are no more serious than may be expected in a coin-tossing experiment.[1]

Plotting the Best-fitting Normal Curve.—We could now use the expected frequencies as the basis of plotting the best-fitting, smooth, normal distribution curve for the memory-test data. If plotting such a curve is our only objective, however, we have done some unnecessary work. A shorter procedure for locating enough points for drawing the smooth best-fitting curve will now be explained. It follows precisely the same principles laid down in the previous discussion. But instead of being tied down to class intervals and their midpoints for our x values, we instead arbitrarily choose standard scores at convenient values $.5\sigma$ apart, as in the first column of Table 7.2. Since they are simple numbers, no interpolation will be necessary in using Table B. Since the positive standard scores duplicate the negative ones, half the work of looking up y values is obviated, unless one wishes to repeat the process as a check. The expected frequencies are again found by multiplying y by iN/σ, in this case, by 40. As before, this step is for the sake of obtaining frequencies in the proportions comparable with those obtained for a particular N (86), a particular σ (6.45), and a particular size of class interval (3).

The frequencies found in this manner will not correspond to midpoints of class intervals, however, but to other score-point positions on the scale. These points will be $.5\sigma$ apart, starting at the mean and going both ways. They correspond to the z scores given in the first column of Table 7.2. We need to find the corresponding X values for these z values. The first step

[1] The customary way of determining whether the discrepancies between theoretical and obtained frequencies are so large as not to be attributed to sampling errors is to employ the chi-square test (see Ch. 11). The chi-square test, as applied to the normal-curve hypothesis, enables us to arrive at a decision as to the probability that an obtained set of frequencies is not normally distributed.

TABLE 7.2.—OBTAINING THE BEST-FITTING NORMAL CURVE FOR THE DATA ON THE MEMORY TEST FOR THE PURPOSE OF PLOTTING THE CURVE

(1)	(2)	(3)	(4)	(5)
z Standard score	y From Table B	f_e Expected frequency	x Deviation	X Raw score
+3.0	.0044	0.2	+19.4	45.5
+2.5	.0175	0.7	+16.1	42.2
+2.0	.0540	2.2	+12.9	39.0
+1.5	.1295	5.2	+ 9.7	35.8
+1.0	.2420	9.7	+ 6.4	32.5
+0.5	.3521	14.1	+ 3.2	29.3
0.0	.3989	16.0	0.0	26.1
−0.5	.3521	14.1	− 3.2	22.9
−1.0	.2420	9.7	− 6.4	19.7
−1.5	.1295	5.2	− 9.7	16.4
−2.0	.0540	2.2	−12.9	13.2
−2.5	.0175	0.7	−16.1	10.0
−3.0	.0044	0.2	−19.4	6.7

The numbers in the columns are obtained as follows:
Column (1): Arbitrarily chosen.
Column (3): $40 \times y$.
Column (4): $6.45 \times z$.
Column (5): $x + 26.1$.

is to find the corresponding x deviations by the formula

$$x = z\sigma \quad \text{(A deviation derived from a standard score)} \quad (7.4)$$

These are shown in column (4) of Table 7.2. The X points corresponding to x deviations can be found by the formula

$$X = M + x \quad \text{(A measurement estimated from a deviation)} \quad (7.5)$$

which, in this problem is $X = 26.1 + x$. The X values we want are shown in the last column of Table 7.2.

Having these score points and their corresponding frequencies, we can construct the graph shown in Fig. 7.3. The observed frequencies (f_o) are also plotted as circlets to show where they fall with respect to the best-fitting normal curve. The reasonableness of the fit is rather obvious. It would probably have been not so easy to duplicate this normal curve by the smoothing process recommended in Ch. 3. We may say by way

of general conclusion that if our obtained mean and standard deviation approximate closely the mean and sigma of the population from which our sample came, and if the distribution for the population is normal, it looks like the curve in Fig. 7.3.

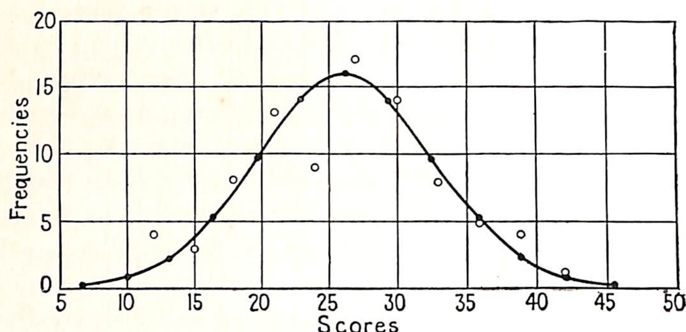


FIG. 7.3.—The best-fitting normal distribution curve for the memory-test data. Obtained frequencies are represented by circlets. The normal curve is "best-fitting" in the sense that it has the same mean and standard deviation as the obtained distribution.

AREAS UNDER THE NORMAL CURVE

Perhaps the greatest usefulness of the normal curve lies in the relationship of the amount of area under the curve lying between certain limits on the base line. In terms of mental-test scores, for example, this simply means the number or percentage of the cases to be expected between two score points. This is because the area under the curve represents the number or percentage of cases. The total area is equal to N, the *total number* of cases. But if we think in terms of a standard curve where $N = 100$, we can readily deal with percentages. For example, 50 per cent of the surface lies above the mean and 50 per cent below. We can also think in terms of a standard curve whose total surface is equal to 1, or unity. In this instance we deal with proportions. The proportion of the area, or cases, lying above the mean is .5 and the proportion below is .5. The statistical tables are given in terms of a total area of 1, and the areas of certain segments are listed as proportions, but it is just as easy to talk in terms of percentages. A percentage is a proportion multiplied by 100, and a proportion is a percentage divided by 100. Thus .46 of the surface is 46 per cent; and 72 per cent of the cases is .72 of the surface, etc.

Proportion of the Area between the Mean and Some Measurement or Score.—We have already had occasion to say that the interval extending one standard deviation on either side of the mean includes about two-thirds of the cases. To say the same thing in another way, from the mean to plus 1σ are to be expected about one-third of the cases, and from

the mean to minus 1σ, another one-third of the cases. We can verify this by referring to Table B and looking up the proportion of the area between the mean and 1σ (*i.e.*, a z equal to 1.00). The area given to four decimal places is .3413, or three thousand four hundred thirteen ten-thousandths of the area. If there were a normal distribution with 10,000 cases, 3,413 of them would be expected between the mean and 1σ. In terms of percentage, it would be 34.13 per cent, or 34.13 cases in 100. The total interval from $+1\sigma$ to -1σ contains twice this area, or .6826, or 68.26 per cent. Figure 7.4 illustrates these facts graphically. We now see that this is a little more than two-thirds (which would be 66.67 per cent), but with small deviations from normality occurring on every hand we can afford to be so rough with our expectations as to give it as two-thirds.

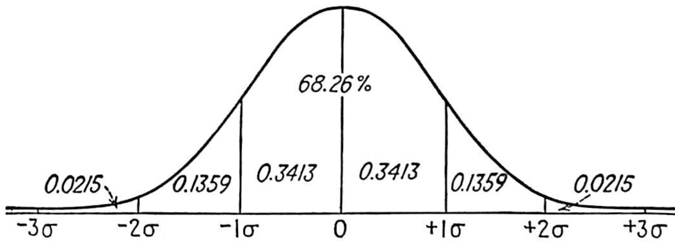


FIG. 7.4.—Different percentages of area under the normal curve within the various 1-sigma units on the base line.

From Table B, we can also see that between the mean and a point 2σ distant (either above or below, *i.e.*, either $+2\sigma$ or -2σ), we should expect .4772 of the total surface, or 47.72 per cent of the cases. Included in the range from -2σ to $+2\sigma$, we should find twice this proportion, or .9544 of the area, or 95.44 per cent of the cases. Out to 3σ from the mean extends .4987 of the area, and in both directions from the mean to 3σ we find twice this, or .9974 of the area. Only 26 cases in 10,000 (10,000 − 9,974), therefore, should be expected *beyond* the range from -3σ to $+3\sigma$ in a large sample.

To take another example of a less special nature, how much of the area under the normal curve will be found between the mean and $+0.78\sigma$? From the table, we find this to be .2823. In still another problem, how many cases lie between the mean and -1.47σ? From the table, we find this to be .4292. Figure 7.5 illustrates these two cases. It will be seen that the positive or negative sign of z merely tells us whether the area extends above the mean or below. The numerical *size* of z, whether positive or negative, determines the *amount* of area between the mean and the point.

So far we have begun each problem of this type with some particular z or standard measurement. Let us start the problem a step or two further back and begin with some raw score or measurement. In the more practical case, we begin with X, not z. In the memory-test data, we may inquire what proportion of the cases come between the mean (26.1) and a point of 35 on the scale of measurement. This point deviates 8.9 units

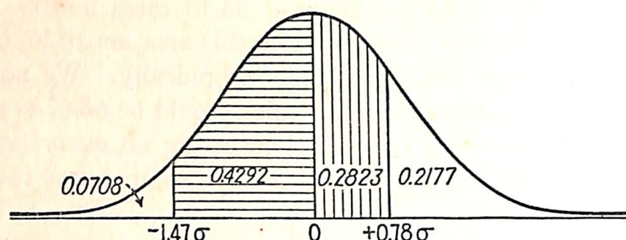


FIG. 7.5.—Proportions of the total area under the normal curve within certain standard-score limits on the base line.

from the mean ($X - M = +8.9$). This is the deviation x. The standard score z is x/σ, which equals $8.9/6.45 = +1.38$. *Everything must be transformed into standard measure before the probability table may be utilized.* Entering the table with a z of 1.38, we find the corresponding area to be .4162. In other words, 41.62 per cent of the cases in a normal distribution would be found between the mean and 35 points on the scale. In the memory-test data, 41.62 per cent of 86 is 35.8, or, in whole numbers,

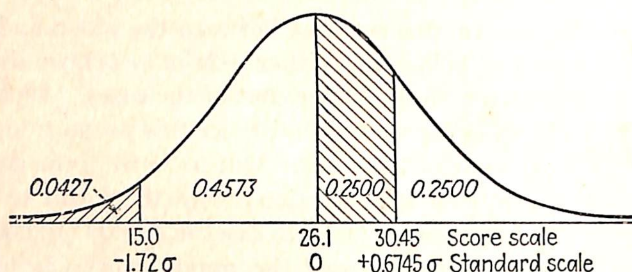


FIG. 7.6.—Proportions of the cases to be expected between certain score limits in the memory-test data, on the assumption that the distribution is normal.

36 cases. In a similar manner, which the student should verify, between the mean and a score of 20 are .3276 of the cases, or approximately 28. Between the mean and 15 are about 39 cases of the 86, and if we go on down to a score point of 5, we find 49.95 per cent of the cases.

Special interest attaches to the question of the proportion of cases between the mean and a score of 30.45. It will be found that the standard score corresponding to this is 0.6745. From the table we find that the proportion of the area to this point is .25, or exactly one-fourth.

This case is illustrated in Fig. 7.6. In short, the point at 0.6745σ corresponds to a distance of $1Q$ from the mean.

The Area above or below a Certain Point on the Scale.—For a given deviate or standard score, Table B also gives us the proportion of the areas above a certain point on the scale or below it. Above a point at $+1\sigma$ will be found .1587 of the area. This is found in column (*C*) of Table B, because when a vertical line is erected at $+1\sigma$ (see Fig. 7.7), it divides the total area under the curve into two portions, the one above the line being the smaller of the two. Below the point $+1\sigma$ is the remainder of the area, or the larger portion [found in column (*B*) of the table], including .8413, or 84.13 per cent of the area. If we were interested in the point -1σ, the larger portion under the curve is now above the point of division and is found in column (*B*), whereas the portion

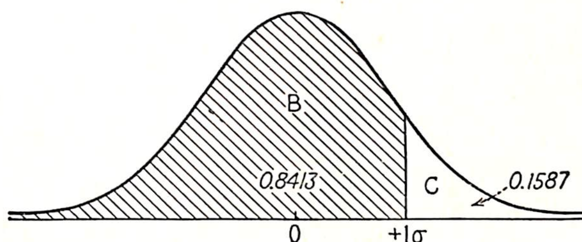


FIG. 7.7.—Proportions of the area above and below the standard score of $+1\sigma$ and under the normal curve.

below, being the smaller of the two, is found in column (*C*). The situation is just reversed to the case where the division comes at $+1\sigma$. It is necessary to keep in mind in this kind of problem whether the area we wish to know is under the smaller end of the curve, all on one side of the mean, or whether it is under the larger side of the curve extending across the mean.

The proportion of the area above the point at $+0.78\sigma$ is in the smaller portion, and found in column (*C*), it is .2177. The area below -1.47σ is also under the smaller portion of the curve, and from column (*C*), we find that it is .0708 (see Fig. 7.5). The area *above* the point -1.47σ would be equal to $1.0 - .0708$, which is .9292. Or it can be found from column (*B*), since it occupies the larger portion under the curve, and this also gives us .9292. Or, from Fig. 7.5, we can see that it is the sum of the area from the point to the mean (.4292) plus .500, which gives the same result.

In the memory-test data, where the mean is 26.1 and σ is 6.45, we may ask for the percentage of the cases to be expected below a score of 15. The deviation from the mean is 11.1. When this is divided by 6.45, we find that the z-score is -1.72. Corresponding to a z of -1.72 is an area

of .0427 in the tail of the normal curve (see Fig. 7.6). We may expect 4.27 per cent of the cases below a score of 15; or, out of 86, this would be 3.7 cases. Above a score of 15, we should expect the remainder of the cases, naturally; *i.e.*, a proportion of .9573, a percentage of 95.27, and in number of cases, 82.3. Above a score of 30.45, which corresponds to a z score of $+0.6745$, we should expect 25 per cent of the cases.

Area between Two Points on the Scale.—The first case of this kind of problem has already been mentioned when we asked for the proportion of the area between -1σ and $+1\sigma$ and the like. When the two score points are on two sides of the mean, it is simply a matter of summing the two areas between the mean and the two points. For example, between the points -1.47σ and $+0.78\sigma$, we have the two areas .4292 and .2823 to add (see Fig. 7.5). The result is .7115, or 71.15 per cent.

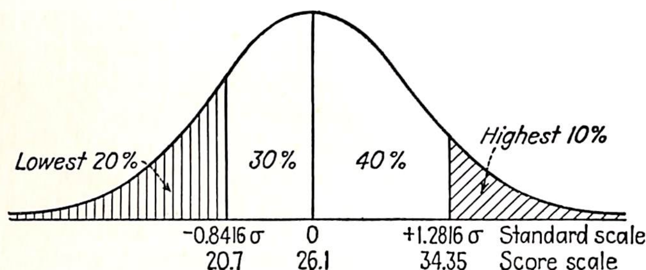


Fig. 7.8.—Score points above or below which certain percentages of the cases are expected in the memory-test distribution, assuming normality of distribution.

When the two points lie on the same side of the mean, it is a matter of subtracting the smaller area from the larger, more inclusive area. For example, the area between points at $+1\sigma$ and $+2\sigma$ can be found by first obtaining from the table the area from the mean to $+1\sigma$ (which is .3413) and the area from the mean to $+2\sigma$ (which is .4772). The area we seek is $.4772 - .3413 = .1359$ (see Fig. 7.4). The area between points -2σ and -3σ would be the area .4987 [from Table B, column (A)] minus .4772 (from the same source). The difference is equal to .0215, which is illustrated in Fig. 7.4.

The area between two raw-score points again involves the determination of z scores as the first step. In the memory-test data, between scores 10 and 20, which correspond to z scores of -2.50 and -0.945, respectively, the area is the difference between .4938 and .3276, which is .1662, or 16.62 per cent. The areas from the mean to the two z scores are found as usual in Table B. As one more example from the same data, the proportion of the cases between scores of 30 and 35 is equal to .1888, for the z scores are $+0.605$ and $+1.38$, respectively, and the area to the

mean in the two cases .2274 and .4162. The student should verify these estimates.

Points above or below Which Certain Proportions of the Cases Fall.—The next problems reverse the processes that have just been described. Before, we were given points on the scale of measurement to determine areas; now we are given areas from which to determine points on the scale. For example, above what point in the normal curve does the highest 10 per cent of the cases come? Ten per cent is a proportion of .10. We could now use Table B in reverse, but it is much more convenient to utilize Table C, which gives the proportions in even steps. We are faced with a problem that gives the proportion in the tail of the curve; so we look in the last column for *C*, the smaller area. We find the *z* score corresponding to it to be 1.2816. This will be with plus sign, since we are talking about the highest 10 per cent (see Fig. 7.8). Had we asked below what point does the *lowest* 10 per cent fall, the answer would have been -1.2816σ. If the question is, "Above what score lies the highest 80 per cent of the cases?" we are then dealing with the larger proportion under the curve; so we look for the proportion of .80 in the first column of Table C. The corresponding *z* score is -0.8416σ (see Fig. 7.8). Had we asked for the point below which is the *lowest* 80 per cent, the answer would have been +0.8416.

To apply these same questions to the memory-test data, we need go a step further and transform the *z* scores into terms of the raw-score scale. The highest 10 per cent come above a *z* of +1.2816. Multiplying this by σ (which is 6.45), we obtain the deviation (x) of +8.27. The mean (or 26.1) plus 8.27 gives us a score of 34.37 points. The highest 10 per cent in a normal curve with mean of 26.1 and sigma of 6.45 would come above the point 34.37. It happens that this point comes close to the division point between two class intervals, or 34.5. In the actual distribution (see Table 7.1), 10 cases, or close to 12 per cent, were scores of 35 or above, which is good agreement. Ten per cent would have called for 8.6 cases, or 9 in whole numbers.

The highest 80 per cent of the cases, which we found to come above a *z* score of -0.8416σ, will be expected above a raw score of what? The deviation of this point from the mean is −5.43 points, or a score of 20.67. This comes close to another division point between class intervals, namely, 20.5. In the actual distribution, 71, or 82.5 per cent, of the cases are above a score of 20.5. Again the agreement between obtained proportion and expected proportion is quite close. To take one more case, which gives a point exactly between class intervals, we ask above what point are 93.2 per cent of the cases? The point turns out to be a score of

16.5 points (the student should verify this). The actual percentage of cases above this score point is 92—again a very close agreement.

Centiles and Corresponding z Scores.—By now it may be apparent that we can look up in the tables the z score corresponding to any given centile. For example, p_{90} is the point below which are 90 per cent of the cases. Entering Table C with .90 in column (*B*), we find the corresponding z to be $+1.2816$. Corresponding to p_{80} is the z score of $+0.8416$. We could find the corresponding raw-score points corresponding to all these z scores for any particular distribution. If the assumption of normal distribution is valid, this procedure would be an advance step over the recommendation of smoothed ogives for setting up centile norms. But if there is any noticeable skewing in the distribution, this procedure would be rather questionable. The smoothed-ogive method would leave the actual skewness taken into account. Since further measurements with the same test will probably yield the same kind of distribution from the same population, this deviation from normality should be represented in the norms.

It can now be explained how, earlier (see Table 6.5), we arrived at the spacing of centile scores on the profile chart (Fig. 6.5). The values given to represent the spacing of the centiles are the z scores corresponding to them, and they were obtained as was explained in the preceding paragraph. The result is to normalize the distribution of all tests, whether the original measuring scale gave a normal distribution or not. There is, in other words, a general underlying assumption of normal distribution of the population in all the abilities represented in the profile chart. The most important gain in so doing is to transform measurements of all abilities into the terms of a common intelligible scale.

The Points between Which Lie Certain Proportions of the Middle Cases.—Among the problems involving area under the curve, there remains the case in which, given the area of a central group, what are the score limits of that group? The only practical case here occurs when the central group is evenly balanced on either side of the mean; the middle 50 per cent, 80 per cent, or 90 per cent. Those groups, it will be remembered, are significant in connection with indicators of variability and are given distinction in the graphic device illustrated in Fig. 6.7. Here, however, we are talking about the best-fitting normal curve and not the original distribution. The middle 50 per cent extends from Q_1 to Q_3, or from p_{25} to p_{75}. Going to the tables with a proportion of .75, we find the corresponding z to be, as we should expect, 0.6745σ. The two points bounding this middle 50 per cent are -0.6745 and $+0.6745$. In the dis-

tribution of memory-test scores, these points would correspond to actual scores of 21.75 and 30.45. The interpolated Q_1 and Q_3 in this same obtained distribution were 21.00 and 30.85, respectively, or not very far from those estimated in the best-fitting curve. The middle 80 per cent extends from p_{10} to p_{90}. We have previously determined these to be at a distance of 1.2816σ, minus and plus. The corresponding raw scores are 17.83 and 34.37. The interpolated 10th and 90th centiles are 17.1 and 35.3, again in close agreement. This kind of problem has really little application in psychological and educational statistics, but is included for the sake of completeness and with the hope that it may lend further insight into the several ramifications of the normal distribution curve. All other problems having to do with area illustrated above do have numerous and valuable applications, some of which we shall meet in Ch. 12.

Exercises

1. *a.* Toss six pennies 64 times. After each throw, note and record the number of heads. Compare your obtained frequencies with the expected frequencies. Plot frequency polygons of the two distributions. Compute the mean and standard deviation of the distribution.
   b. Toss the same six pennies 64 times more, obtaining a new set of data like the first. Compute the mean and standard deviation of this distribution, and make comparisons with the first obtained distribution and with the theoretical distribution.
   c. Combine the two distributions into a single one. Are the frequencies now any nearer the expected ones? Compute the mean and standard deviation. Are they any nearer the mean and standard deviation of the theoretical distribution?
   d. One more experiment may be tried in which some of the outcomes with a small number of heads are not counted, but another throw is immediately substituted. Every second case in which at a glance you can tell the number of heads is small, should be ignored and the trial repeated. Again, obtain 64 record trials. This situation illustrates a biased sampling. What is the effect upon the frequencies?
   e. What would happen in another set of trials if one penny were left head up, only the remaining five being thrown each time but all six coins being observed and all heads being counted?

2. Determine the standard scores for all the midpoints in the distribution of Data 7*A*. Also determine the z scores for the following raw scores: 40, 55, 72, 85, 95.

3. From Table B, determine the ordinate value at each midpoint of distribution 7*A*.

4. Find the expected frequency for each class interval, and tabulate them and the observed frequencies in parallel columns. State some inferences that you can draw from your results.

5. Find the best-fitting normal curve for Data 7*A* after the manner of Table 7.2. Plot the curve along with the obtained frequencies.

6. Find the proportions and percentages of the areas under the normal curve between the mean and the following z scores: −2.15 −1.85 −0.19 +0.375 +1.1 +3.52.

DATA 7*A*.—DISTRIBUTION OF SPELLING-TEST SCORES IN A SUPERIOR GROUP OF FRESHMEN*

Scores	*f*
82–85	1
78–81	8
74–77	8
70–73	5
66–69	34
62–65	21
58–61	39
54–57	32
50–53	20
46–49	7
42–45	3
38–41	0
34–37	1
Sum.......	179
Mean.....	61.1
σ..........	8.4

* The test was one of the Cooperative series, and the scores are T scores (see Ch. 12).

7. Find the proportions and numbers of the cases to be expected between the mean and the following scores in Data 7*A*: 35 45 60 75 79.5 38.35.

8. Find the proportions of the area *above* the following z scores: +2.15 +1.62 +0.175 −0.36 −1.9 −2.8. Also, *below* the following z scores: −3.85 −1.225 −0.6745 +0.005 +1.75 +2.3.

9. Find the proportions and numbers of cases to be expected in Distribution 7*A* *above* the following scores: 80 55 65 27 69.5 54.5 41.5. And *below* these scores: 85 45 56 35 77.5 41.5 61.5. Whenever possible, compare expected with obtained frequencies.

10. Find the proportions of the area falling *between* z scores of −1.50 to +1.25 −0.05 to +2.76 +0.55 to +0.95 −2.78 to −1.12 +3.15 to +2.95 −0.72 to −1.05 +1.24 to −0.33

11. Find the proportions and numbers of cases to be expected in Distribution 7*A* between scores of 70 and 80 35 and 45 45 and 65 65.5 and 77.5 49.5 and 57.5 45.5 and 65.5 65.5 and 69.6 61.5 and 65.6 53.5 and 57.5. Whenever possible, compare expected with obtained frequencies.

12. Give in terms of standard measurements the points *above* which the following percentages of the cases fall in the normal distribution: 85, 55, 35, 42.3, 66.7, and 9.42 per cent.

13. Give the z scores *below* which the following proportions of the cases will fall: .14 .62 .375 .418 .729

14. *Above* what scores in Distribution 7*A* will the following percentages of the cases be expected: 12, 54, 84.13, 5.75, and 68.4 per cent?

15. *Below* what scores in Distribution 7*A* should we expect the following number of cases: 11 63 89.5 123 162? Compare expected with actual cumulative frequencies.

16. What z scores correspond to the following centile ranks: 75 62.5 16.7 5 99?

17. Between what score limits in Distribution 7*A* should we expect the middle 80 per cent of the cases? The middle 50 per cent? The middle 90 per cent? Compare these with the interpolated limits for these same percentages.

CHAPTER 8

CORRELATION

No single statistical procedure has opened up so many new avenues of discovery in psychology and education as that of correlation. This is understandable when we remember that scientific progress depends upon finding out what things are co-related and what things are not. A *coefficient of correlation* is a single number that tells us to what extent two things are related; to what extent variations in the one go with variations in the other. Without the knowledge of how one thing varies with another, we should find predictions impossible. And wherever causal relationships are involved, without knowledge of covariation, we should be unable to control one thing by manipulating another.

For example, when we know that the higher a girl's score in a clerical-aptitude test the higher the average performance she is likely to exhibit after training, we can thereafter use scores on this test to predict level of proficiency. We say that there is a high positive correlation between aptitude-test score and clerical success. We discover this fact by finding a coefficient of correlation between scores of a number of girls and measures of clerical performance later for the very same girls. We can never compute a coefficient of correlation on one person alone, nor can we compute it without having made two sets of measurements on the same individuals, or on matched pairs of individuals. In this instance, if we consider that the aptitude test has measured individual differences in some quality or qualities that lead to success, *i.e.*, in the sense of a cause of clerical success, then we can not only predict future success for individuals but also promote high general efficiency in any group of clerks by selecting those with high scores. Thus are studies leading to prediction and control of human affairs promoted because correlation techniques are available. Without some device like this for checking up on a test, we have only vague notions concerning its effectiveness, unless, indeed, its effectiveness is so obvious to direct observations as to require no inspection by correlation methods, which is highly unlikely.

THE MEANING OF CORRELATION

Some Examples of Correlation between Two Variables.—The coefficient of correlation is one of those summarizing numbers, like a mean

or a standard deviation, which, though it is a single number, tells a story. It can vary from a value of +1.00, which means perfect positive correlation, through zero, which means complete independence or no correlation whatever, on down to −1.00, which means perfect negative correlation.

A Case of Perfect, Positive Correlation. Figure 8.1 illustrates an instance of perfect positive correlation. It is a fictitious case, for such exact agreement between two things is rarely or never experienced, certainly not in psychology or education. Here we have assumed two tests, X and Y. Ten individuals have received scores in the two tests. The pairs of scores are as follows:

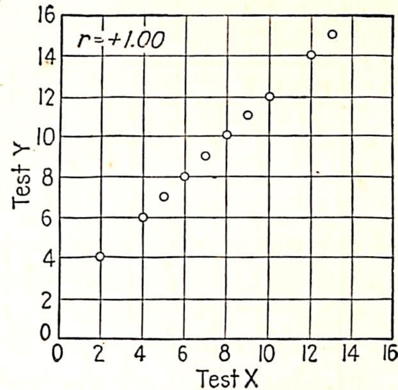


FIG. 8.1.—A simple correlation chart showing the kind of relationship between X and Y scores when the correlation is +1.00.

Individual	*A*	*B*	*C*	*D*	*E*	*F*	*G*	*H*	*I*	*J*
Score in test X	2	4	5	6	7	8	9	10	12	13
Score in test Y	4	6	7	8	9	10	11	12	14	15

Looking down the rows of scores, each pair made by one individual, we readily conclude that each person's score in Y is two points higher than his score in X. In terms of a simple equation, $Y = X + 2$. There are *no exceptions*, which makes the correlation perfect.

To take another instance:

Individual	*A*	*B*	*C*	*D*	*E*	*F*	*G*	*H*	*I*	*J*
Score in test P	1	3	4	5	7	8	9	11	12	15
Score in test Q	2	6	8	10	14	16	18	22	24	30

In this situation, each person's score in Q is two times that in P, again without exception; there is perfect agreement, and the coefficient of correlation would be +1.00. The equation for predicting Q from P is $Q = 2P$.

A Case of High Positive Correlation.—In Fig. 8.2, we have illustrated a case of correlation that is positive but less than +1.00. The graphic

picture of the individuals shows that, in general, a person who is high in test X is also high in test Y, and one who is low in X is also likely to be low in Y. The actual scores for these 10 people are listed in the first two columns of Table 8.1. It will be seen that although the individuals are arranged in rank order for scores in X, there are some deviations from this rank order when we inspect their scores in Y. The coefficient of correlation by computation is equal to +.76. We shall soon see how this was obtained but first simply note by comparison of Figs. 8.1 and 8.2 how the individuals are scattered in the diagrams. In Fig. 8.1, they line up in perfect file from lowest to highest. In Fig. 8.2, they tend to fan out or to diverge from a strict line-up, but a definite trend of relationship can be observed. The amount of spreading in Fig. 8.2 as compared with that in Fig. 8.1 (in which it is, of course, none) illustrates the difference between correlations of +1.00 and +.76.

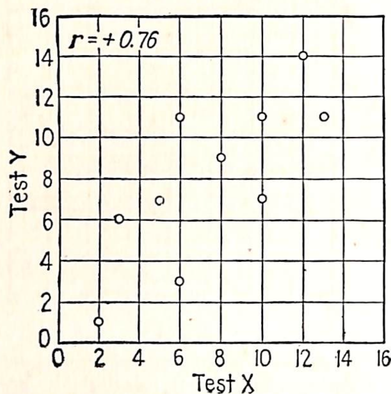


FIG. 8.2.—A correlation chart illustrating the kind of situation when the correlation is +.76.

A Case of Low Positive Correlation. A third instance is shown in Fig. 8.3, in which the spreading effect to which our attention was called before is even greater. The coefficient of correlation here is +.14; in other words, close to zero. This being true, a person with high score in X is likely to be almost anywhere, within the total range, in terms of his Y score. The three highest people in X, with scores of 10, 12, and 13, scatter all the way from 3 to 11 in test Y. The three lowest people in test X, with scores of 1, 3, and 4, scatter all the way from 2 to 9 in test Y. Although there is a trace of relationship between X scores and Y scores, it is very weak. The actual scores may be compared in Table 8.3.

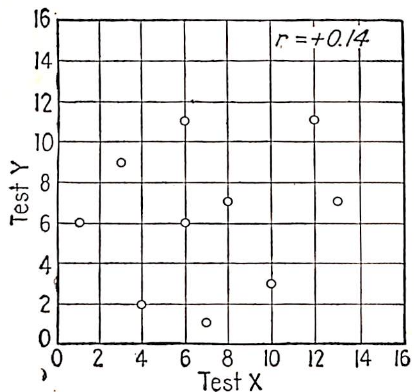


FIG. 8.3.—An example of a correlation chart when the correlation is only +.14.

A Case of High Negative Correlation.—The situation that obtains when there is a negative correlation is shown in Fig. 8.4. Here the coefficient is −.69. Compare this diagram with that in Fig. 8.2, and it will be appar-

ent that the trend of the points is along the other diagonal now, from upper left to lower right. This illustrates the fact that persons making high scores in X are likely to make low scores in Y, and persons making low scores in X are likely to make high scores in Y. This inverse *order* of relationship is also apparent in the actual scores in the first two columns of Table 8.2. The numerical *size* of the coefficient (.69) is nearly the same as for the correlation in Fig. 8.2 (.76). It will be seen that the width of scatter of the points is about the same in the two cases. A perfect negative correlation would be pictured as a line of dots like that in Fig. 8.1 but it would slant downward instead of upward from left to right. The algebraic sign of the coefficient of correlation therefore merely has to do with the *direction* of the relationship between two things, whether direct or inverse, and the size of the coefficient (distance from zero) has to do with the *strength* or *closeness* of the relationship.

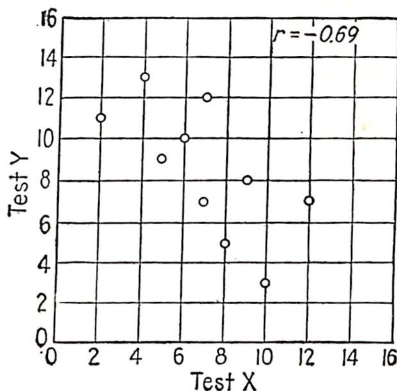


FIG. 8.4.—An example of a correlation chart when the correlation is −.69.

How to Compute a Coefficient of Correlation

The Product-moment Coefficient of Correlation.—The standard kind of coefficient of correlation and the one most commonly computed is Pearson's product-moment coefficient. The basic formula is

$$r_{xy} = \frac{\Sigma xy}{N\sigma_x\sigma_y} \quad \text{(Basic formula for a Pearson product-moment coefficient of correlation)} \tag{8.1}$$

where r_{xy} = correlation between X and Y.
x = deviation of any X score from the mean in test X.
y = deviation of the corresponding Y score from the mean in test Y.
Σxy = sum of all the products of deviations, each x deviation times its corresponding y deviation.
σ_x and σ_y = standard deviations of the distributions of X and Y scores.

The steps necessary are illustrated in Table 8.1. They will be enumerated here:

Step 1. List in parallel columns the paired X and Y scores, making sure that corresponding scores are together.

TABLE 8.1.—CORRELATION BETWEEN TWO SETS OF MEASUREMENTS OF THE SAME INDIVIDUALS; UNGROUPED DATA; PRODUCT-MOMENT COEFFICIENT OF CORRELATION

	X	Y	x	y	x^2	y^2	xy
	13	11	+5.5	+3	30.25	9	+16.5
	12	14	+4.5	+6	20.25	36	+27.0
	10	11	+2.5	+3	6.25	9	+ 7.5
	10	7	+2.5	−1	6.25	1	− 2.5
	8	9	+0.5	+1	0.25	1	+ 0.5
	6	11	−1.5	+3	2.25	9	− 4.5
	6	3	−1.5	−5	2.25	25	+ 7.5
	5	7	−2.5	−1	6.25	1	+ 2.5
	3	6	−4.5	−2	20.25	4	+ 9.0
	2	1	−5.5	−7	30.25	49	+38.5
Sums	75	80	0.0	0	124.50	144	102.0
Means	7.5	8.0			Σx^2	Σy^2	Σxy

$$\sigma_x = \sqrt{\frac{124.50}{10}} = \sqrt{12.450} = 3.528$$

$$\sigma_y = \sqrt{\frac{144}{10}} = \sqrt{14.4} = 3.795-$$

$$r_{xy} = \frac{\Sigma xy}{N\sigma_x\sigma_y} = \frac{102.0}{(10)(3.53)(3.79)} = \frac{102.0}{133.90} = +.76$$

An alternative solution without computing the σ's:

$$r_{xy} = \frac{\Sigma xy}{\sqrt{(\Sigma x^2)(\Sigma y^2)}} = \frac{102.0}{\sqrt{(124.5)(144)}} = \frac{102.0}{\sqrt{17{,}928.0}} = \frac{102.0}{133.90} = +.76$$

Step 2. Determine the two means M_x and M_y. In Table 8.1, these are 7.5 and 8.0, respectively.

Step 3. Determine for every pair of scores the two deviations x and y. Check them by finding algebraic sums, which should be zero.

Step 4. Square all the deviations, and list in two columns. This is for the purpose of computing σ_x and σ_y.

Step 5. Sum the squares of the deviations to obtain Σx^2 and Σy^2.

Step 6. From these values compute σ_x and σ_y.

Step 7. For every person, find his xy product (last column of Table 8.1). Sum these for Σxy.

Step 8. We are now ready for formula (8.1). In the illustrative problem, the arithmetic is given following Table 8.1.

A Shorter Solution.—There is an alternative and shorter route that omits the computation of σ_x and σ_y, should they not be needed for any other purpose. The formula is

$$r_{xy} = \frac{\Sigma xy}{\sqrt{(\Sigma x^2)(\Sigma y^2)}} \qquad \text{(Alternative formula for a Pearson } r\text{)} \qquad (8.2)$$

The solution with this formula is also given with Table 8.1, and it leads to the same coefficient. In both cases, two significant digits have been saved in r, for the reason that for so small a number of cases the sampling error in r is so relatively large that more than two digits would be rather deceiving as to accuracy. When N is large—200 or more—three-place accuracy in r may more properly be reported.

Computing a Negative Coefficient.—As another example of the computation of r, when the correlation is *negative*, Table 8.2 is presented. The operations are just the same, step by step. The only thing new is the care that must be taken with algebraic signs.

TABLE 8.2.—A NEGATIVE CORRELATION IN UNGROUPED DATA BY THE PRODUCT-MOMENT METHOD

	X	Y	x	y	x^2	y^2	xy
	12	7	+5	−1.5	25	2.25	− 7.5
	10	3	+3	−5.5	9	30.25	−16.5
	9	8	+2	−0.5	4	.25	− 1.0
	8	5	+1	−3.5	1	12.25	− 3.5
	7	7	0	−1.5	0	2.25	0.0
	7	12	0	+3.5	0	12.25	0.0
	6	10	−1	+1.5	1	2.25	− 1.5
	5	9	−2	+0.5	4	.25	− 1.0
	4	13	−3	+4.5	9	20.25	−13.5
	2	11	−5	+2.5	25	6.25	−12.5
Sums	70	85	0	0.0	78	88.50	−57.0
Mean	7.0	8.5			Σx^2	Σy^2	Σxy

$$\sigma_x = \sqrt{\frac{78}{10}} = \sqrt{7.8} = 2.79$$

$$\sigma_y = \sqrt{\frac{88.5}{10}} = \sqrt{8.85} = 2.97$$

$$r_{xy} = \frac{-57.0}{(10)(2.79)(2.97)} = \frac{-57.0}{82.863}$$
$$= -.69$$

Computing r from Original Measurements.—In both examples thus far, we have been dealing with a small number of observations and ungrouped data. When the data are more numerous, we resort to grouping into class intervals; but first let us see another procedure with ungrouped data, which does not require the use of deviations. It deals entirely with original scores. When raw scores are small numbers or when a good calculating machine is available, this is the best procedure. The formula may look forbidding but is really easy to apply:

$$r_{xy} = \frac{N\Sigma XY - (\Sigma X)(\Sigma Y)}{\sqrt{[N\Sigma X^2 - (\Sigma X)^2][N\Sigma Y^2 - (\Sigma Y)^2]}} \quad \text{(A Pearson } r \text{ computed from original data)} \qquad (8.3)$$

where X and Y are original scores in variables X and Y. Other symbols tell what is done with them. We follow the steps that are illustrated in Table 8.3.

Step 1. Square all X and Y measurements.
Step 2. Find the XY product for every pair of scores.

TABLE 8.3.—CORRELATION OF UNGROUPED DATA COMPUTED FROM THE ORIGINAL MEASUREMENTS

X	Y	X^2	Y^2	XY
13	7	169	49	91
12	11	144	121	132
10	3	100	9	30
8	7	64	49	56
7	2	49	4	14
6	12	36	144	72
6	6	36	36	36
4	2	16	4	8
3	9	9	81	27
1	6	1	36	6
Sums 70	65	624	533	472
ΣX	ΣY	ΣX^2	ΣY^2	ΣXY

$$r^2_{xy} = \frac{[N\Sigma XY - (\Sigma X)(\Sigma Y)]^2}{[N\Sigma X^2 - (\Sigma X)^2][N\Sigma Y^2 - (\Sigma Y)^2]}$$
$$= \frac{(4{,}720 - 4{,}550)^2}{(6{,}240 - 4{,}900)(5{,}330 - 4{,}225)}$$
$$= \frac{(170)^2}{(1{,}340)(1{,}105)}$$
$$= \frac{28{,}900}{1{,}480{,}700}$$
$$= .019518$$
$$r_{xy} = \sqrt{.019518}$$
$$= +.14$$

Step 3. Sum the X's, the Y's, the X^2's, the Y^2's, and the XY's.
Step 4. Apply formula (8.3).

The author has found it more convenient, particularly when machine work can be done, to compute r^2_{xy} first by the formula

$$r^2_{xy} = \frac{[N\Sigma XY - (\Sigma X)(\Sigma Y)]^2}{[N\Sigma X^2 - (\Sigma X)^2][N\Sigma Y^2 - (\Sigma Y)^2]} \tag{8.4}$$

and then finally extract the square root to find r_{xy}, as shown just below Table 8.3.

Preparing a Scatter Diagram.—When N is large, even when N is moderate in size, and when no calculating machine is available, the customary procedure is to group data in both X and Y and to form a scatter diagram or correlation diagram. The choice of size of class interval and limits of intervals follows much the same rules as were given in Ch. 3. For the sake of a clearer illustration of the procedure, a smaller number of classes will be employed in the problem now to be described. The data were scores earned by a class in educational measurements in two objectively scored examinations, one of which stressed statistical methods and the other of which stressed tests and measurements.

In setting up a double grouping of data, a table is prepared with columns and rows—columns for the dispersions of Y scores within each class interval for the X scale, and rows for the dispersions of X scores within each class interval for the Y scale. Along the top of the table (see Table 8.4) are listed the score limits for the class intervals in test X. Along the left-hand margin are listed the score limits for the class intervals in test Y. We make one tally mark for each individual's X and Y scores. For example, if one individual had a score of 83 in test X and a score of 121 in test Y, we place a tally mark for him in the *cell* of the diagram at the intersection of the column for interval 80–84 in X and the row for interval 120–124 in Y. All other individuals are similarly located in their proper cells.

TABLE 8.4.—A SCATTER DIAGRAM OF THE SCORES IN TWO ACHIEVEMENT TESTS

X: Scores in First Achievement Test

Y: Scores in Second Achievement Test	60-64	65-69	70-74	75-79	80-84	85-89	90-94	95-99	f_y
135-139								/ 1	1
130-134				/ 1	/ 1		/ 1		3
125-129				/ 1		// 2	/ 1		4
120-124			/ 1	//// 4	//// 4	////// 6	// 2		17
115-119			/////// 7	///// 5	/////// 7	// 2	/ 1		22
110-114	/ 1	//// 4	// 2	///////// 9	//// 4	// 2			22
105-109	/ 1	/ 1	// 2	///// 5	/ 1				10
100-104	/ 1	/// 3		/ 1	/ 1				6
95-99		// 2							2
f_x	3	10	12	26	18	12	5	1	87 N

When the tallying is completed, we write the number of cases, or the *cell frequency*, in each of the cells. Next we sum the cell frequencies in the rows separately, recording each frequency in the last column under the heading f_y. When this column is filled, we have the total frequency

distribution for test Y. We also sum the cell frequencies in all the columns, writing them in the bottom row with its heading f_x. When completed, this row gives us the total frequency distribution for test X. We can check the summing of the cell frequencies by adding up the last row and last column. Their sums should, of course, both equal N, in this case, 87. The check does not, however, guarantee correct tallying. This can be checked partly when we correlate either test with another one and compare total frequency distributions or when we have knowledge of the correct frequency distribution of Y or of X from any other source. There are times when it is wise to do the entire tallying two times and to compare all cell frequencies in the two attempts. It is very easy to place a tally mark in the wrong cell.

Computing the Pearson r from a Scatter Diagram.—When the product-moment r is computed from a scatter diagram, the formula becomes

$$r_{xy} = \frac{\frac{\Sigma x'y'}{N} - (c'_x c'_y)}{(\sigma'_x)(\sigma'_y)} \qquad \text{(Pearson } r \text{ from grouped data)} \qquad (8.5)$$

where x' and y' = deviations from the guessed mean in terms of the class interval as the unit.

c'_x and c'_y = corrections in X and Y in class-interval units.

σ'_x and σ'_y = standard deviations in X and Y in terms of the class interval as the unit.

The details of application of this equation will now be explained and illustrated.

Determining the Corrections and Standard Deviations.—The procedures for calculating the corrections and standard deviations for both X and Y separately are no different than was previously described in Ch. 5. From Table 8.5 we have the necessary information, which is used as follows:

$$c'_x = \frac{\Sigma fx'}{N} = \frac{20}{87} = .230$$

$$c'_y = \frac{\Sigma fy'}{N} = \frac{-30}{87} = -.345$$

$$\sigma'_x = \sqrt{\frac{\Sigma fx'^2}{N} - (c'_x)^2} = \sqrt{\frac{206}{87} - .0529} = \sqrt{2.3149} = 1.52$$

$$\sigma'_y = \sqrt{\frac{\Sigma fy'^2}{N} - (c'_y)^2} = \sqrt{\frac{224}{87} - .1190} = \sqrt{2.4557} = 1.57$$

Determining the Sum of the Cross Products.—The new process to be mastered here is the calculation of the cross products, or products of the

TABLE 8.5.—SCATTER DIAGRAM FOR COMPUTING A PEARSON r

X: Examination in Statistics

Y: Examination in Educational Measurements

	60–64	65–69	70–74	75–79	80–84	85–89	90–94	95–99	f_y	y'	fy'	fy'^2	$\Sigma x'y'$ +	$\Sigma x'y'$ −
135–139								^{16} 1 $_{16}$	1	+4	+4	16	16	
130–134				1	^{3} 1 $_{3}$	^{6}	^{9} 1 $_{9}$		3	+3	+9	27	12	
125–129				1	^{2}	^{4} 2 $_{8}$	^{6} 1 $_{6}$		4	+2	+8	16	14	
120–124			$^{-1}$ 1 $_{-1}$	4	^{1} 4 $_{4}$	^{2} 6 $_{12}$	^{3} 2 $_{6}$		17	+1	+17	17	22	1
115–119			7	5	7	2	1		22	0	0	0	0	0
110–114	^{3} 1 $_{3}$	^{2} 4 $_{8}$	^{1} 2 $_{2}$	9	$^{-1}$ 4 $_{-4}$	$^{-2}$ 2 $_{-4}$			22	−1	−22	22	13	8
105–109	^{6} 1 $_{6}$	^{4} 1 $_{4}$	^{2} 2 $_{4}$	5	$^{-2}$ 1 $_{-2}$				10	−2	−20	40	14	2
100–104	^{9} 1 $_{9}$	^{6} 3 $_{18}$		1	$^{-3}$ 1 $_{-3}$				6	−3	−18	54	27	3
95–99		^{8} 2 $_{16}$							2	−4	−8	32	16	
f_x	3	10	12	26	18	12	5	1	87		−30	224	134	−14
x'	−3	−2	−1	0	+1	+2	+3	+4			$\Sigma fy'$	$\Sigma fy'^2$		
fx'	−9	−20	−12	0	+18	+24	+15	+4	+20	$=\Sigma fx'$				
fx'^2	27	40	12	0	18	48	45	16	206	$=\Sigma fx'^2$				
$\Sigma x'y'$ +	18	46	6	0	7	20	21	16	134			$\Sigma x'y'=$	+120	
$\Sigma x'y'$ −			1	0	9	4			−14					

moments, and their sum, in other words, $\Sigma x'y'$. It is best to begin with the idea that every cell has its own $x'y'$ product and to keep that idea in mind. In fact, it is well to determine the $x'y'$ product for every cell in which individuals fall and to write it in as was done in Table 8.5.

The $x'y'$ product for any cell is simply the product of the x' value times the y' value of that cell, close watch being kept of algebraic signs. This matter is easily checked, of course, by making sure that the sign of every $x'y'$ product is positive in the upper right quarter of the chart and also the lower left quarter, but they are all negative in the upper left and lower right quarters. This rule presupposes that the X measurements are increasing from left to right and that the Y measurements are increasing from below upward.

Having given every cell its $x'y'$ value and having recorded it in the upper left-hand corner of the cell, we next note how many individuals have that $x'y'$ value—in other words, the frequency in that cell. We multiply the cell product by the frequency, and in Table 8.5 these products are recorded with algebraic sign in the lower right-hand corners of the cells. All that remains now is to summate them. We do this both in the columns and in the rows for the sake of checking, for this is an unusually critical number in the correlation formula, and because of the many steps involved in deriving it there are many opportunities for errors. The last two columns in Table 8.5 are devoted to the sums of $fx'y'$ values in the rows. We keep the sums of the positive products in one of these

columns and the sums of the negative products in the other. The last two rows of the table are reserved likewise for summing the positive and negative sums in the columns. Summing everything in the last two columns (also in the last two rows) of the table gives us $\Sigma x'y'$, and the two estimates should check exactly. For the illustrative problem, the positive sum is 134 and the negative is -14, leaving a net positive sum $\Sigma x'y'$ of 120. We now have everything we need for calculating r. Applying formula (8.5), we have

$$r_{xy} = \frac{\frac{120}{87} - (.23)(-.345)}{(1.52)(1.57)}$$

$$= \frac{1.3793 + .0794}{2.3864}$$

$$= \frac{1.4587}{2.3864}$$

$$= .61$$

Interpretations of a Coefficient of Correlation

How High Is Any Given Coefficient of Correlation?—Any coefficient of correlation that is not zero and that is also statistically significant denotes some degree of relationship between two variables.[1] But we need further orientation on the matter, for the strength of relationship can be regarded from a number of points of view, and it is not correct from any one of these points of view to say that the degree of relationship is exactly proportional to r. The coefficient of correlation does *not* give directly anything like a percentage of relationship. We cannot say that an r of .50 indicates two times the relationship that is indicated by an r of .25. Nor can we say that an increase in correlation from $r = .40$ to $r = .60$ is equivalent to an increase in correlation from $r = .70$ to .90. The coefficient of correlation is an index number, not a measurement on a linear scale of equal units.

A General Verbal Description of Coefficients.—Our interpretation of the size of r depends very much upon what we propose to do with it or the reasons why we computed it. What would be a large correlation coefficient for one purpose would be regarded as a small one for another. Interpretation is therefore largely a relative matter; relative to the area of investigation in which we are working and to other factors. But taking correlations just at large, without particular regard to their use

[1] For a treatment of the topic of statistical significance of a coefficient of correlation, see Ch. 9.

and as a general orientation, we may say that the strength of relationship can be described roughly as follows for various r's:

Less than .20 . . . . Slight; almost negligible relationship
.20–.40 . . . . Low correlation; definite but small relationship
.40–.70 . . . . Moderate correlation; substantial relationship
.70–.90 . . . . High correlation; marked relationship
.90–1.00 . . . . . . Very high correlation; very dependable relationship

It should be said that the coefficients should be interpreted as stated only when, by comparison with the standard error of r, they prove to be significant. It should also be said that the same interpretations apply alike to negative and positive r's of the same numerical size. An r of $-.60$ indicates just as close a relationship as an r of $+.60$.

Particular Uses Have a Bearing on Interpretation of r.—The general descriptive list just given should be qualified by making references to particular uses of r. One common use is to indicate the agreement of scores on an aptitude test with measures of scholastic or of vocational success. Such a correlation is known as a *validity coefficient.* It is an index of the practical validity of a test. Chapter 18 will deal extensively with this subject. Common experience shows that the validity coefficient for a single test may be expected within the range from .00 to .60, with most of them in the lower half of that range. Validity coefficients for composite scores based upon combinations of several different kinds of tests are likely to be distinctly higher, ranging up to .80 in rare instances but hardly ever above the latter figure. Many who have employed tests for vocational guidance or vocational selection have followed a tradition which may be credited to C. L. Hull[1] some twenty years ago, that the minimum validity coefficient for a test of practical usefulness is about .45. Recent experiences have shown that this standard is too rigid and that there are many considerations other than validity which determine the usefulness of a test in any given situation, as will be shown in Ch. 15.

It is well recognized that a *reliability coefficient,* which in very general terms is a correlation of a test with itself, is usually a much higher figure than a validity coefficient. Following the leadership of T. L. Kelley,[2] there has been a general tradition that to be sufficiently reliable for discriminating between individuals, a test should have a reliability coefficient of at least .94. Some have been more liberal in this regard, allowing a

[1] Hull, C. L. Aptitude testing. Yonkers-on-Hudson: World, 1928. Ch. 8.

[2] Kelley, T. L. Interpretation of educational measurements. Yonkers-on-Hudson: World, 1927. Pp. 210*ff*.

minimum of .90, while others have been more demanding, with a requirement of a minimum of .96. These standards are rarely attainable, and it is safe to say that most tests in use fail to meet them. As a matter of fact, there are many very useful tests whose reliability coefficients are in the .80's and even below. It is coming to be recognized that validity is much more important than reliability, and, in fact, it is possible for a test to be sufficiently valid for practical purposes without being very reliable. Tests with reliability coefficients as low as .35 have been found useful when utilized in batteries with other tests.[1] Such tests have been known with validities as high as .35. They could theoretically have validities much higher than that. Reliability and validity depend upon many considerations that we cannot go into here. These problems will be treated in Chs. 17 and 18. It is sufficient to say that one must be a relativist when dealing with problems of test reliability and validity. The student's interpretation of a coefficient of correlation, like his interpretation of other statistics, is subject to considerable revision as he knows more about its uses. While these qualifications mentioned regarding reliability and validity need to be made, the fact remains that in practice we expect reliability coefficients to be in the upper brackets of r values, usually .80 to .98, and validity coefficients to be in the lower brackets, usually .00 to .80.

When one is investigating a purely theoretical problem, even very small correlations, if statistically significant (undoubtedly not zero), are often very indicative of a psychological law. Whenever a relationship between two variables is established beyond reasonable doubt, the fact that the correlation coefficient is small may merely mean that the measurement situation is contaminated by many things uncontrolled or not held constant. One can readily conceive of an experimental situation in which, if all irrelevant factors had been held constant, the r might have been 1.00 rather than .20. For example, the correlation between an ability score and scholarship is .50, since both are measured in a population whose scholarship is also allowed to be determined by effort, attitudes, marking peculiarities of the instructors, and what not. Were all the other determiners of scholarship held constant and were both aptitude and marks perfectly measured, the r would be 1.00 rather than .50. This line of reasoning indicates that where any correlation between two things is established at all, and particularly where there is a causal relationship involved, the fundamental law implies a perfect relationship. Thus, in nature, correlations of zero or 1.00 are the rule between variables when

[1] Guilford, J. P. New standards for test evaluation. *Educ. & Psychol. Meas.*, 1946, **6**, 427–438.

isolated. The fact that we obtain anything else is because of the inextricable interplay of variables that we cannot measure in isolation.

The practical conclusion from this is that *a correlation is always relative to the situation under which it is obtained, and its size does not represent any absolute natural or cosmic fact.* To speak of *the* correlation between intelligence and scholarship is absurd. One needs to say *which* intelligence, measured under *what* circumstances, in *what* population, and to say *what kind* of scholarship, measured by *what* instruments, or judged by *what* standards. *Always, the coefficient of correlation is purely relative to the circumstances under which it was obtained and should be interpreted in the light of those circumstances; very rarely, certainly, in any absolute sense.*

How much faith one should place in any relationship shown by a coefficient of correlation also depends upon the urgency of the outcome. There are probably many medical treatments, such as some inoculations, vaccines, and the like, concerning which the knowledge is rather incomplete, which are administered even though the correlation between the treatment and living (or between nontreatment and dying) is of the order of .10 to .20. Although the probabilities of living may be increased by only 1 per cent by the treatment, the saving of 1 life in 100 is regarded as worth the effort. If a procedure in education promised only 1 per cent improvement over guesswork, we should pay little attention to it, because the seriousness of the outcome would not justify the means. It may be said in passing, however, that failures to predict in vocational and educational practice are more generally recognized by reason of correlational checkup than are failures to predict in medical practice, where correlational checkup is less often made. In addition to the difference in relative seriousness of the outcomes of prescription in the two cases, this factor of better knowledge of goodness of results may be an important reason for the higher standards of prescriptive accuracy demanded in education than are sometimes required in other fields.

Graphic Representations of Correlations

In presenting the facts of correlation to the layman, who is probably not accustomed to thinking in terms of numerical indices in any case and who has probably never learned of the coefficient of correlation, it is better to convey the idea of a relationship in other ways, preferably in the form of a diagram of some kind. Figure 8.5 and 8.6 are two examples of how this might be done. Figure 8.5 is a bar diagram showing for each level of aptitude score, on a nine-point scale (stanine scale), the percentage of pilot students who graduated from flying schools. The actual percentages are given for those who are interested in simple numbers. In

spite of the unusually large samples the percentages are given to two significant digits only. The number of students in each stanine group is given for those who have some appreciation of the stability offered by large samples.

The other diagram, Fig. 8.6, shows the average rating of flying proficiency made by cadets at each stanine level, and only the average. Some investigators connect successive pairs of points with lines, but in

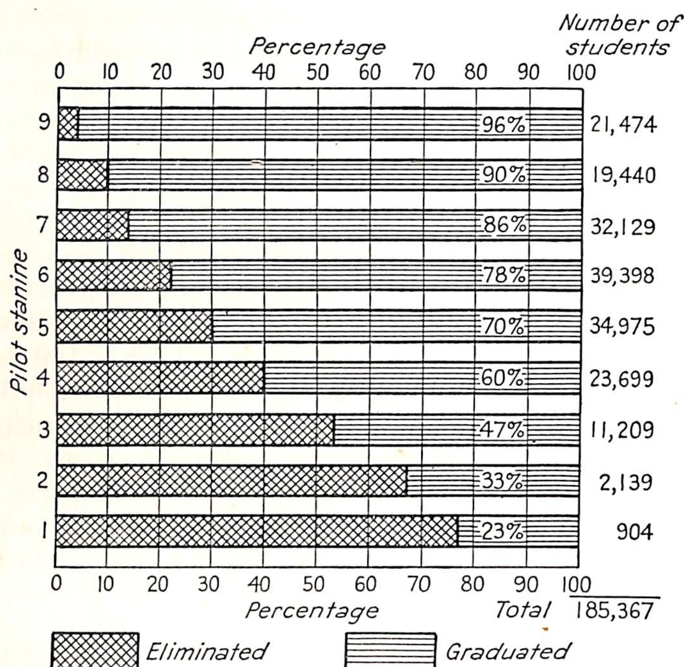


FIG. 8.5.—Correlation between the pilot-aptitude score (pilot stanine) and the criterion of graduation-elimination from flying training in the AAF illustrated by a bar diagram. (*Based upon Stanines: Selection and classification for air crew duty. Washington D.C.: Headquarters, Army Air Forces, 1946.*)

this particular instance the linear trend is so clear that a straight line has been drawn by inspection to fit the trend. It is assumed that minor deviations that occur are due to sampling errors. A warning should be given in connection with this type of figure. It can give an impression of degree of correlation far in excess of that justified. Not shown are the widths of dispersions of individuals, at different stanine levels, in this case. While the averages of columns do not deviate much from a straight line, many individual cases may deviate considerably. There are ways of representing average discrepancies of individuals from such a regression

line (see Ch. 15) which could be used to give the reader some idea of their seriousness.

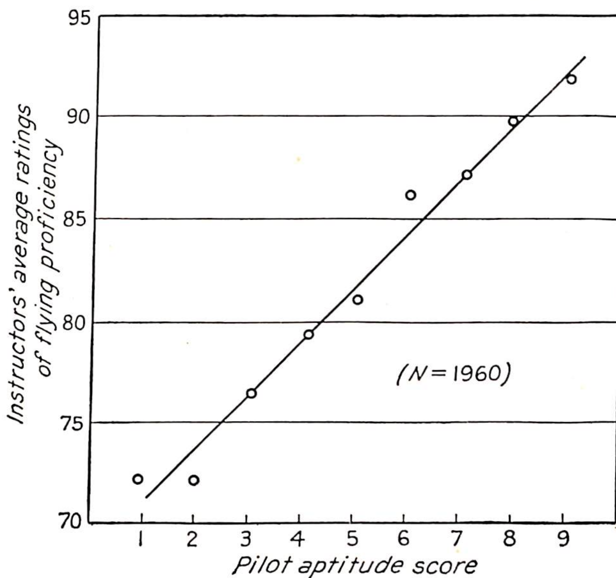


Fig. 8.6.—Correlation between pilot-aptitude scores and instructors' ratings of flying proficiency illustrated by means of a regression line that is based upon the averages of ratings for different aptitude-score levels.

Assumptions Underlying the Product-moment Correlation

The student should be warned before leaving this chapter concerning the restrictions that should be observed in the use of the Pearson coefficient of correlation. The most important requirement for the legitimate use of the Pearson r is that the trend of relationship between Y and X be rectilinear, in other words, a straight-line regression. This can be determined, as a rule, by inspection of the scatter diagram. If the distribution of the cases within the correlation diagram appears to be elliptical, without any indications of a decided bending of the ellipse, the chances are that the relationship is rectilinear. Even if it is not, the deviation from a straight-line relationship may be so slight that we may assume rectilinearity as a first approximation, and the degree of correlation indicated by r will be fairly close to any index of correlation like the *correlation ratio* (see Ch. 13) that is applied when there is curvature in the trend. When there is an obvious bending of the distribution of cases, a correlation ratio, or some other special coefficient, is indicated as the best index of correlation.

There are in educational and psychological measurements certain factors that produce artificially curved scatters in the correlation diagram. This may happen when one or both distributions taken alone are badly skewed and the skewing is produced artificially by the faulty measuring scale, with its systematically shifting unit of measurement. If there is good reason to believe that this may be the case, one solution would be to normalize the skewed distribution by methods described in Ch. 12. When distributions are corrected for skewness, the curvature in the regression is frequently eliminated, and linearity is then obtained. If curvature still remains, then the Pearson r is not to be used to indicate the amount of correlation.

There is nothing in what has been said to demand that the Pearson r is to be computed only with normal distributions. The forms of distributions may be various, so long as they are fairly symmetrical and unimodal; even rectangular ones would do. The important consideration is whether in all columns the dispersions are approximately equal, as indicated by the column standard deviations, and also in all rows. This condition goes by the name *homoscedasticity*. When columns (and rows) are relatively homoscedastic, we may compute a Pearson r. This condition will prevail generally when the two distributions are fairly symmetrical within themselves; so we need not go so far as to compute standard deviations of columns and rows in order to find out.[1] It is when distributions are markedly skewed that significant departures from homoscedasticity occur.

Figure 8.7 is presented to show graphically the kind of scatter plots one might expect when one or both distributions are symmetrical or skewed. In each diagram the form of distribution assumed is shown along the X or Y dimension. In diagram A both distributions are assumed to be normal. The probable scatter of the cases within the square area is elliptical. The contour of the ellipse (and of corresponding objects in the other diagrams) is not drawn so as to enclose *all* the cases but to include the central mass of them. The regression in diagram A is clearly rectilinear, and homoscedasticity prevails. In diagram B, X is normally distributed and Y is negatively skewed. The trend of the cases is definitely curved and the distribution is not homoscedastic in either vertical or horizontal arrays ("array" is a general term including both rows and columns). In diagram C, with skewing in the same direction in both X and Y distributions, the regression appears to be rectilinear but the dispersion is not homoscedastic. In diagram D, the skewing is in

[1] Some writers suggest that only when both distributions are normal or nearly so will the conditions be fully satisfied for computating a Pearson r. In practice probably no one insists upon normal distributions.

opposite directions and there is neither rectilinearity nor homoscedasticity. Only in the case of diagram A would one justifiably compute a Pearson product-moment coefficient of correlation. In a later chapter (Ch. 13) other types of coefficients of correlation will be described which might be

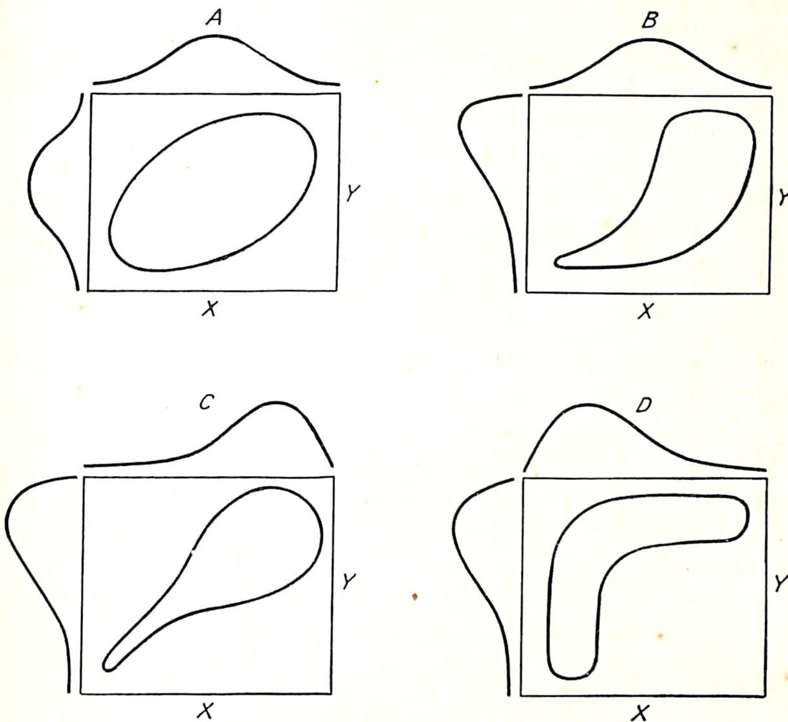


FIG. 8.7.—Hypothetical forms of scatter plots in a correlation diagram when the forms of distribution of X and Y values differ. Diagram A shows linear regression and homoscedasticity; B and D show curved regression and lack of homoscedasticity; and C shows linear regression but lack of homoscedasticity.

applied to the data in diagrams B, C, and D if one could justify the appropriate assumptions that must be made.

Exercises

1. Using the first 10 pairs of scores in the list in Data 8A, compute a Pearson r between any two parts that you or your instructor selects. Use formulas 8.1 and 8.2. Find a similar coefficient, using the last 10 pairs of scores in the same two variables. State your conclusions.

2. Correlate the first 10 pairs of scores for any two other parts, using formulas 8.3 and 8.4. Correlate the same two parts, using the last 10 pairs and the same formulas. State your conclusions.

3. Prepare a scatter diagram for the correlation of Parts III and IV, or any other two parts, including all 40 cases. Compute a Pearson r using formula (8.5). State conclusions.

DATA 8*A*.—SCORES EARNED BY FORTY HIGH-SCHOOL STUDENTS IN SEVEN PARTS OF THE GUILFORD-ZIMMERMAN APTITUDE SURVEY*

Part I Verbal Comprehension	Part II Reasoning	Part III Numerical Operations	Part IV Perceptual Speed	Part V Spatial Orientation	Part VI Spatial Visualization	Part VII Mechanical Knowledge
22	11	24	29	27	39	30
8	5	22	40	16	23	21
19	6	44	36	14	12	21
32	8	72	32	21	20	33
13	2	25	46	25	20	29
24	5	30	47	2	6	8
22	4	38	49	15	37	35
35	1	54	53	34	28	16
18	7	37	51	37	46	30
13	10	61	50	38	46	35
53	23	56	45	22	41	38
15	9	42	48	18	5	18
34	18	30	25	40	58	46
15	2	42	48	12	21	17
27	4	28	28	31	26	24
19	9	32	40	11	13	19
29	4	24	37	26	0	27
24	9	42	58	21	21	23
27	9	54	54	23	20	30
16	5	42	44	29	24	34
56	12	67	48	20	40	26
22	5	58	48	28	41	20
32	4	57	33	20	4	16
18	8	49	47	19	36	42
24	15	87	52	36	44	26
22	12	14	48	25	16	27
22	10	38	46	21	0	20
21	21	32	33	11	43	37
13	10	52	40	29	35	11
23	3	60	49	43	13	37
2	10	29	49	10	21	27
20	4	50	55	22	8	27
25	11	76	43	26	20	26
14	6	40	38	35	8	46
11	2	32	56	38	4	26
2	9	61	45	20	10	20
38	17	56	67	25	20	35
16	6	61	42	29	23	21
14	4	17	44	26	7	21
23	25	61	48	23	29	16

* Part I is a vocabulary test; Part II is composed of arithmetic reasoning problems; Part III is composed of simple number operations; Part IV is on matching visual objects differing very little; Part V involves awareness of spatial relationships; Part VI requires imagination of an object turned in space; and Part VII is on common knowledge of tools and their use, automobile parts and functions, and common trade knowledge. The intercorrelations in this particular sample will be found to be generally low except between Parts I and II and between V and VI.

4. Do the same as in Exercise 3 for one or more other pairs of variables in Data 8*A*. How many pairs of coefficients of correlation are possible with Data 8*A*? State a general rule for the number of intercorrelations when there are *n* variables.

5. Compute the Pearson *r* for Data 8*B*. Interpret your findings.

6. Find five Pearson *r* coefficients reported in the literature. Tell what variables were being correlated in each case. Interpret the results. Are the coefficients about the sizes you would have expected for the things correlated? Were there any special conditions that may have biased the amount of correlation in one way or another?

DATA 8*B*.—A SCATTER DIAGRAM OF REACTION-TIME MEASUREMENTS AND GRADES EARNED IN GENERAL PSYCHOLOGY

Reaction time to auditory stimulus	Grades in psychology								
	55–59	60–64	65–69	70–74	75–79	80–84	85–89	90–94	95–99
.180–.189					1				
.170–.179						1			
.160–.169				2	1	1		1	
.150–.159					1	1	1		
.140–.149	1			1	2	1	1	1	
.130–.139	1			6	2	6	1	3	
.120–.129			1		2	3	3		1
.110–.119				2	1	2			
.100–.109								1	

CHAPTER 9

THE RELIABILITY AND SIGNIFICANCE OF STATISTICS

In this chapter we raise the very important question as to how near the "truth" are statistical answers such as means, standard deviations, proportions, and the like. As was said before, any measured sample is usually employed to represent a larger population. A population, from the statistical point of view, is any arbitrarily defined group. The term will be more fully explained in later paragraphs.

Our sampling has to be limited for practical reasons; we cannot measure total populations, or at least it is generally inefficient and unnecessary to do so. Yet we usually wish to generalize beyond our sample, arriving at scientific decisions that transcend the observations made at a particular time and in a particular place, or reaching administrative decisions that apply to larger groups of individuals. In preceding chapters we have been concerned with *descriptive statistics* only. The computed values were used to describe the properties of particular samples. If we want to apply those same descriptive statistics beyond the limits of samples, we must know how much risk of being wrong we take. In general terms, the statistics stressed in this chapter are designed to do that very thing. They are known as *sampling statistics.*

To be more specific, when we obtain the mean of a sample that is measured in some respect, before we say that this obtained mean also describes the central tendency of the population sampled, we need to find some basis for believing that it does not deviate very far from the population mean. Fortunately, there is a statistical procedure that will inform us about how far our obtained mean probably deviates from the population mean, provided certain conditions, to be explained later, have been satisfied. The statistic that will do this is known as the *standard error of the mean.* In a similar manner, there are standard errors of other sample statistics—medians, standard deviations, proportions, correlation coefficients, and the like—which inform us of the accuracy of our obtained figures as estimates of the corresponding population values.[1]

[1] In some statistical writings a population mean is referred to as a *true* mean, and other statistics likewise called *true* when reference is made to population values. It seems better practice to steer clear of philosophical issues by avoiding reference to *truth.* Some

Some Principles of Sampling

Before going into the treatment of sampling statistics, it is necessary to have clearly in mind the essential facts about the process of sampling. The application of sampling statistics depends upon certain *conditions* of sampling. If these are not satisfied, standard errors, no matter how accurately computed, may give wrong impressions. At best, they give us only estimates from which we can make decisions and draw conclusions, never with complete conviction but with various degrees of assurance. After making this frank concession to the limitations of sampling statistics, it should also be asserted that without them we can hardly draw any generalized conclusions at all that would be of scientific or practical value.

Populations and Samples.—It is time that we had a better definition of *population*. Some statisticians call it *universe*. In any case, the statistician's idea of population is quite different from the popular idea. Rarely would any statistical study regard the entire population of a nation, a city, or of some geographical region as its *universe*. The population in a statistical investigation is always arbitrarily defined by naming its unique properties. It might be the entering freshman class in a certain university, or the part of the freshman class entering a certain college or even a certain course. It might be the male sixteen-year-olds in a given school district; the children of Mexican parentage in a certain city; or the registered democratic voters in the New England states. All of these examples are of groups of human individuals. Populations could, of course, be defined as species, or phyla, or order of animals or of plants. There are also populations of observations or of reactions of a certain kind—simple reactions to sound stimuli, word-association reactions, judgments of pleasantness of colors, and the like, from the psychological laboratory. It is probably the nonhuman groups that have seemed to require the more general term *universe* as an alternative to the more restricted term *population*. In this volume we shall use the term *population* in the broad sense to include all sets of individuals, objects, or reactions that can be described as having a unique pattern of qualities.

The reader will also find the term *population* used in this chapter in a more restricted and technical sense. Two samples may be said to come from the *same* population if it can be shown that they are alike in just one respect, *e.g.*, alike in *IQ*, in score on a certain memory test, or in

writers also define a population mean as the mean of a very large number of means of samples taken at random from the same population. This sort of mean would presumably be numerically almost identical with the actual population mean.

showing a like reaction to some proposal. The *likeness* is usually defined statistically in terms of equal means or proportions and equal variances or dispersions. It ordinarily takes rigorous statistical tests to satisfy the scientific investigator that two or more samples have come from the same population. The *same* population, here, then, is same perhaps in only one respect, ignoring differences in other respects.

Parameters and Statistics.—If we were to measure all the individuals of a population and actually to compute the indices of central tendency, dispersion, and correlation, as we ordinarily do for samples, we would obtain what the statisticians call *parameters*. The population parameters

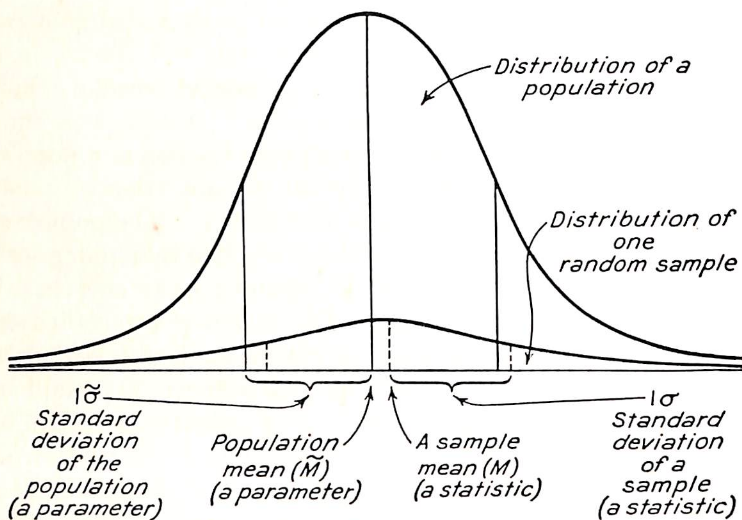


FIG. 9.1.—A comparison of a population distribution and a sample distribution, also of population parameters and sample statistics.

exist whether we compute them or not, if we ignore the dynamic changes that may be occurring and assume for practical purposes that these parameter values are fixed at least for a time.

Figure 9.1 illustrates the distinction between population parameters and sample statistics. The larger distribution is that of the entire population. The smaller distribution is of a sample drawn at random from that population. The population parameters, mean and standard deviation, are symbolized by $\tilde{M}$ and $\tilde{\sigma}$, each with a tilde over it. It will be noted that in this particular sample the mean (M) and the standard deviation (σ) do not coincide exactly in size with their corresponding parameters ($\tilde{M}$ and $\tilde{\sigma}$). This is characteristic. A second sample would be expected to have still different M and σ, but also similar to $\tilde{M}$ and $\tilde{\sigma}$ in

size. The same sort of parallel could be illustrated with respect to proportions ($\tilde{p}$ and p), semi-interquartile ranges ($\tilde{Q}$ and Q), and coefficients of correlation ($\tilde{r}$ and r). By careful and adequate sampling we hope to arrive at statistics that will approximate the corresponding parameters very closely. By means of standard errors and other sampling statistics, to be discussed later, we estimate how far our obtained statistics may have deviated from their corresponding parameters.

Random Sampling.—It should be kept in mind that the use of sampling statistics (standard errors, and the like) rests on the assumption that the sampling has been random. The best definition of random sampling is that it is selection of cases from the population in such a manner that *every individual in the population has an equal chance of being chosen.* This calls to mind a well-conducted lottery, selective-service numbers, coin tossing, throwing dice, and other operations which allow the "laws of chance" to operate freely.

There are several ways of favoring random sampling from populations. For a population of individuals, if all members are arranged in alphabetical order and one wishes to draw one person in every hundred, the first case might be taken by blind pointing within the first hundred names and every hundredth one following in the list automatically chosen. Tables of random numbers have been published as an aid in random sampling.[1] The numbers themselves have been placed in sequence by some kind of lottery procedure. If individuals in a population are numbered in sequence and thus identified by number, selections can be made by following the random numbers in any systematic way. A random sample should be fairly representative of the population, though in any particular sample, if it is a small one, in particular, by chance it may not be so representative as we would like.

Biased Sampling.—In a biased sample there is a systematic error. Certain types of cases have an advantage over others in being selected. The likelihood of individuals being chosen differs from one to another. A common example of this in educational research is the voluntary return of questionnaires. The names of those who are to receive the questionnaires may, to be sure, be randomly chosen from a much larger group. But suppose that only 60 per cent of those circularized return the questionnaires, which is not an atypical event. The 60 per cent who do return the data might possibly be representative, but there is a strong presumption that in the decision to return or not to return the instru-

[1] Examples are Tippett, L. H. C., Random sampling numbers, Cambridge: Cambridge University Press, 1927, and Lindquist, E. F., Statistical analysis in educational research, Boston: Houghton, 1940, Table 18.

ment there is room for biasing forces to work. Those forces may or may not be relevant to the content of the questionnaire itself. But if the information requested implies favorable or unfavorable facts about the respondant, his associates, or his work, it is quite natural to expect those with a "good" showing will be more inclined to reply than those with a "bad" showing. If the trait of cooperativeness or of responsibility or of dependability of the respondant is involved in the data or even correlated with something wanted in the data, there is also a strong likelihood of bias.

A colossal example of biased sampling is that of the Literary Digest public-opinion poll during the 1936 presidential campaign. Several million post-card ballots were said to have been circulated, certainly anticipating a sample of most generous size. But the mailing lists were made up from telephone directories and automobile registration lists. It so happened that in the poll the telephone subscribers and car owners voted with a majority in favor of the candidate who lost, while the nontelephone subscribers and noncar owners voted at the polls in a more decisive way for the successful candidate. Among those who received post-card ballots there was also probably a selection as to which ones would be most likely to take the trouble to return the card. Those who were most discontented with things as they were and wanted a change most would take the trouble to register a protest straw vote. Those who were contented or who felt somewhat secure as to the outcome would be less likely to return the card. This would also tend to make the vote appear to favor the losing candidate, who was running against an incumbent.

The scientific investigator must be eternally vigilant to the possibility of biased sampling. A good, systematic control of experimental conditions is designed to prevent biased samples or to make known their effects. Where there is less than customary experimental control of the observations, every possible effort should be made to know the conditions under which the data are obtained. Thorough knowledge of the conditions should be a basis for deciding whether selection of cases has been biased. Knowledge of conditions is also essential for the sake of accurate definition of the population sampled.

Stratification in Sampling.—One common procedure that is introduced in sampling to help to prevent biases and also to assure a more representative sample is known as stratification. Stratification is a step in the direction of experimental control. It operates with subgroups of more homogeneous composition within the larger population.

A very common example is to be found in public-opinion polling practices. Suppose the issue to be investigated is public attitude toward a certain piece of labor legislation. It is quite likely that people in the

two major political parties would tend to lean in opposite directions on such an issue. It is probable that people of different socioeconomic categories—professional, business, office worker, semiskilled laborer, and unskilled laborer—would react with some systematic differences on the issue. It is possible, though not so likely, that individuals of the two sexes would tend to respond somewhat differently. Other divisions of the population, such as rural versus urban, regional, and educational groups, might also show systematic differences on the issue. In other words, subgroups of the population are considered with respect to any variable that is suspected of correlating appreciably with the variable being studied. It does not matter that some of the variables are themselves intercorrelated unless such an intercorrelation is very high, in which case it would be superfluous to control selection of samples on both of two variables so closely related.

Having decided which variables are important in sampling, the entire population is studied to see what proportions fall into each category, *i.e.*, what proportions are Democrat or Republican; male or female; urban or rural; in each socioeconomic group; and so on. Any sample to be obtained, then, should have proportional representations from all subgroups. Within each defined subpopulation, *e.g.*, a male, professional, Republican, New England group, random sampling may then be carried out. Random selection of cases would also be made within each of the other defined subpopulations in appropriate numbers. The total sampling procedure here described has been called *stratified-random sampling*.

The importance of the proportional-representation principle and its advantage over a purely random sampling can be readily demonstrated. Suppose that 55 per cent of the Republicans and 45 per cent of the Democrats are in favor of a certain labor bill. In the general population let us assume that 60 per cent are registered Democrats and 40 per cent are registered Republicans. In a random sample of 100 voters one would expect in the long run to draw the two party representatives in about the same ratio, 60/40. This would vary from sample to sample, however, even to the extent that the majority could be reversed; for example, it could even be 45/55. In the typical polling sample we would expect a majority of voters against the bill. If the sample should by chance contain a majority of Republicans, however, the majority might favor the bill. If stratification were applied, we would be sure to have in the sample the ratio 60/40, and with this restriction imposed upon the random sampling we should expect the general population sentiment to be more accurately reflected. Thus it can be seen that a stratified-random sample is likely to be more representative of a total population than is a purely random sample.

Purposive Samples.—A *purposive sample* is one arbitrarily selected because there is good evidence that it is very representative of the total population. Experience has shown in public-opinion polling that there are certain states or regions that come close to national opinion time after time. If one is willing to depend upon this experience, one may use the limited population as the source of the sample to use as a "barometer" for the total population. This is a convenient procedure, but it has the disadvantage that much prior information must have been obtained. There is also a risk that conditions may change to the extent that the particular segment of population no longer represents the total or does not represent it on some new issue.

Incidental Samples.—The term *incidental sample* is applied to those samples that are taken because they are the most available.[1] Many a study has been made in psychology with students in classes of beginning psychology as the samples merely because they are most convenient. Results thus obtained can be generalized beyond such groups with considerable risk. Generalizations beyond any sample can be made safely only when we have defined the population that the sample represents in every significant detail. If we know the significant properties of the incidental sample well enough and can show that those properties apply to new individuals, those new individuals may be said to belong to the same population as the members of the sample. By "significant properties" is meant those variables that correlate with the experimental variables involved. They are the kind of properties considered above in connection with stratification of samples. It is unlikely that membership in political party would have much bearing upon the results of certain experiments performed upon sophomores in a beginning psychology course, but such variables as age, education, social background, and the like may definitely be pertinent. Much depends upon the experimental variable under study; whether it is a motor skill or a social attitude, a suggestible reaction or an interest-test score. If incidental samples are employed, the investigator is under scientific obligation to describe the properties of his group in all aspects that he can conceive as being related to the outcome of the investigation.

The Reliability of Averages

The Distribution of Means of Samples.—Suppose that we are dealing with a population whose mean ($\tilde{M}$) is 50.0 and whose standard deviation

[1] Such a sample is often called "accidental." In no real sense is the sample an accident; it was selected. It would be an "accident," of course, if the sample represented usefully a population in which we want to make predictions of parameters.

($\bar{\sigma}$) is 10.0 on the measuring scale we are using. Such a distribution is illustrated by the top diagram in Fig. 9.2. We do not know these population parameters ordinarily, but for the sake of an illustration we will assume that we do know them here.

Sampling Distributions.—Suppose, next, that we proceed to draw random samples, all of equal size, one at a time, from this population. To satisfy the conditions of random sampling in a strictly mathematical sense, we should replace each sample drawn, after noting the value of each of its members, before drawing the next sample. Each individual should have an equal opportunity of being selected in *every* sample. Having lost one sample, the population is different from what it was originally. When the population is very large, as compared with the size of sample, however, we can forget about this *replacement* requirement for practical purposes. In this case, one sample would "hardly be missed"; that is, its loss would change the chance conditions to an inconsequential degree. We will find, later, that when the size of sample is not decidedly smaller than the population, it is possible to make allowance for this fact.

To take a specific example of random sampling, with the same population described above in mind, let the size of sample be 25. The sample mean will not only differ from sample to sample, but will also usually deviate from the population parameter (in this example, the mean of 50.0). If we have a number of such sample means, we may treat them just as if each were a single observation and set up a frequency distribution of them. This is known as a *sampling distribution.* Such a frequency distribution will be close to the normal form. Normality of distribution of single cases in the total population favors normal distribution of means and of other statistics computed from samples drawn from that population. Even when the population distribution departs from normality, however, the distribution of means of samples drawn from it tend to be normally distributed, unless too small. The smaller the sample, the more does the form of distribution of the population affect the form of distribution of the means. The extreme case would be samples of only one case each, in which event we should expect the distribution of means (if means of one observation each have any real meaning) to be of the same form as that of the population.

A knowledge of the form of sampling distribution of a statistic is very important. Our ability to draw conclusions known technically as *statistical inferences* depends upon knowing the form of distribution of sample statistics. Without knowledge of the form of sampling distribution, many a scientific result would remain inconclusive. The reasons for this will be clearer as we go into the subject of interpretation of standard errors.

The Standard Error of a Mean.—At this stage of getting acquainted with sampling distributions, we are most interested in the dispersion of statistics, in this case, the dispersion of sample means. The reason is that the amount of this dispersion gives us the clue as to how far such sample means may be expected to depart from the population mean. If we are to use a sample mean as an estimate of the population mean, any deviation of such a sample mean from the population mean may be regarded as an error of estimation. The standard error of a mean tells us how large these errors of estimation are in any particular sampling situation. The standard error of a mean is a standard deviation of the distribution of sample means. To distinguish such a standard deviation from the more familiar one that applies to dispersions of individual observations, we call it a *standard error*. In later discussions it may be referred to by use of the abbreviation *SE*.

In order actually to compute the standard error of a mean, we need two items of information: the population parameter $\tilde{\sigma}$, and the size of sample N. Since we do not ordinarily know $\tilde{\sigma}$, it would seem that we could but rarely, indeed very rarely, compute this standard error. There are satisfactory ways of estimating it, however, as we shall see later. The formula for computing the standard error of a mean is

$$\tilde{\sigma}_M = \frac{\tilde{\sigma}}{\sqrt{N}} \quad \text{(Standard error of an arithmetic mean computed from a known population parameter)} \qquad (9.1)$$

where $\tilde{\sigma}$ = standard deviation of the population.
N = number of cases in the sample (not the number of means in the distribution of means).

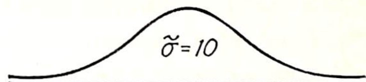
Sample Size and the Standard Error of a Mean.—The standard error of the mean is therefore *directly* proportional to the standard deviation of the population and *inversely* proportional to the size of the sample. More precisely stated, $\tilde{\sigma}_M$ is inversely proportional to the square root of the size of sample. As the individuals of a population scatter more widely, so will the means of samples drawn from that population also scatter more widely. But as we include more individuals in each sample drawn, the *less* widely can the means scatter from their central tendency. In the limiting case, if the sample includes the entire population, the deviation of the sample mean from the population mean can then be only zero, and $\tilde{\sigma}_M$ is zero. In Fig. 9.2 are shown graphically several instances of samples when N varies. The smallest possible sample occurs when $N = 1$. The mean of each sample is then identical with the individual's measurement in that sample. The dispersion of such means is as great as the dispersion of the total population; $\tilde{\sigma}_M$ then equals $\tilde{\sigma}$, which we have

assumed to equal 10. When each sample contains two cases,

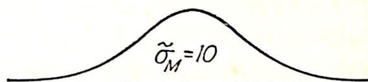
$$\tilde{\sigma}_M = \frac{10}{\sqrt{2}} = 7.07$$

When each sample contains four cases, $\tilde{\sigma}_M = 10/\sqrt{4} = 5$; etc. The remaining cases in Fig. 9.2 should now speak for themselves.

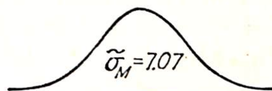
Distribution of individual measures for a whole population



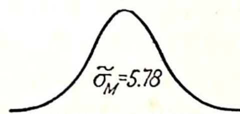
Distribution of means for samples of one case each



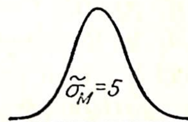
Distribution of means for samples of two cases each



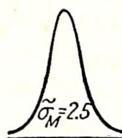
Distribution of means for samples of three cases each



Distribution of means for samples of four cases each



Distribution of means for samples of 16 cases each



Distribution of means for samples of 25 cases each

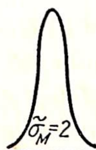


FIG. 9.2.—Showing the hypothetical decrease in variability or fluctuation of the means of samples as we increase the size of the sample drawn at random from a large population. (*Modified from Lindquist, A first course in statistics. Houghton Mifflin, by permission.*)

Estimating the Standard Error of a Mean from σ.—Formula (9.1) requires our knowing the parameter $\tilde{\sigma}$ in order to compute the standard error of a mean. In ordinary practice we must be satisfied with an *estimate* of this standard error. Ordinarily, we have only one sample and its standard

deviation must be utilized as a basis for estimating $\tilde{\sigma}$, and hence for estimation $\tilde{\sigma}_M$. When the sample is known, the formula generally used reads

$$\sigma_M = \frac{\sigma}{\sqrt{N - 1}} \quad \text{(Standard error of a mean estimated from } \sigma\text{)} \quad (9.2)$$

where σ_M = what is ordinarily called the standard error of a mean (as estimated from σ).
σ = standard deviation of the sample.
N = number of cases in the sample.

Strictly speaking, we might well have used the symbol $\tilde{\sigma}_{M_{est}}$ in place of σ_M, to tell the story of its being an estimate of $\tilde{\sigma}_M$. The symbol $\tilde{\sigma}_M$ corresponds to σ_M in much the same way that $\tilde{\sigma}$ corresponds to σ, the one being a parameter and the other an estimate of it given by a sample. That is, σ may be regarded as $\tilde{\sigma}_{est}$. We shall see that it is a biased estimate, but nevertheless an estimate, and we can allow in part for the bias.

Estimating $\tilde{\sigma}$ from a Sample.—The standard deviation in a sample is likely to be smaller than that for the population from which the sample came. Recall from the discussion in an earlier chapter (Ch. 5) that as samples become small the total range of measures is more and more curtailed. This comes about from the fact that extreme deviations in the population are rare and in small samples are likely to be missed. This fact has an effect also upon the size of the standard deviation, though the latter effect is much less drastic than the effect on the range. In small samples, particularly those with N less than 30, the sample σ gives an estimate of the population $\tilde{\sigma}$ that is biased downward. If we want an unbiased estimate of $\tilde{\sigma}$ directly from the sums of squares of a sample, we can use the formula

$$\tilde{\sigma}_{est} = \sqrt{\frac{\Sigma x^2}{N - 1}} \quad \text{(Unbiased estimate of population standard deviation)} \quad (9.3)$$

where Σx^2 = sum of squares in the sample.
N = number of cases in the sample.

Degrees of Freedom.—Formula (9.3) contains an important new concept that will be found liberally utilized hereafter when sampling errors (deviations of statistics from parameters) are mentioned in connection with small samples. Compare formula (9.3) with the basic one for the standard deviation of a sample (formula 5.5), and it will be found that they are identical except for the denominators, which are $(N - 1)$ and N, respectively. The difference between the two may seem very slight (and it *is* slight numerically when N is reasonably large), but there is a very important difference in meaning. In this particular formula, $(N - 1)$ is

known as the number of *degrees of freedom*, which is symbolized by *df*. This is a key concept in recent years in what has been known as *small-sample statistics*. The number of degrees of freedom will not always be $(N - 1)$ but will vary from one statistic to another as will be pointed out in various places later. Let us see why the number is $(N - 1)$ here.

The "freedom" part of the concept means *freedom to vary*. The standard deviation is computed from the variance, and the variance is computed from deviations from the mean. Statisticians often express the matter by saying that one degree of freedom is "used up" when we compute the mean of a sample. This leaves $(N - 1)$ degrees of freedom for estimating the population variance and the standard deviation.

A numerical example will make this clearer. Let us assume five measurements: 5, 7, 10, 12, and 16, the mean of which is 10.0. A mathematical requirement or property of the arithmetic mean is that the sum of the deviations from it equals zero. The five deviations in this sample are -5, -3, 0, $+2$, and $+6$, the sum of which is zero. With this condition satisfied, *i.e.*, the sum equal to zero, how many of these deviations could be simultaneously altered (as if by taking new samplings) and still leave the sum equal to zero? With a little thought or trial and error it will be seen that if any four are arbitrarily changed, the fifth is thereby fixed. We could make the first four -8, -4, $+1$, and -2, which would mean that for the sum to equal zero the fifth has to be $+13$. Try any other changes and if the sum is to remain zero one of the five deviations is automatically determined. Thus only four $(N - 1)$ are "free to vary" within the restriction imposed. The restriction is that the mean is taken as fixed for the sample. In this sense, the computation of the mean "uses up" one degree of freedom. There were N degrees of freedom in computing the mean because the cases were presumably sampled entirely independently. If they were not independently sampled then there were also less than N degrees of freedom in computing the mean. We shall see examples of this later. Freedom means independence and only when there is independence of observations can the "laws of chance" operate freely and the mathematics based upon the "laws of chance" be applied.[1]

The Sample Mean as an Estimate of the Population Mean.—Since the sample σ is a biased estimate of the population $\tilde{\sigma}$, one might expect to hear that the sample M is also a biased estimate of the population $\tilde{M}$. On the other hand, from the discussion in the preceding paragraph, in which it was pointed out that no loss of degrees of freedom affected the computation of the mean, one might conclude that there is no bias in

[1] For an excellent discussion of the general subject of degrees of freedom, see Walker, H. M., Degrees of freedom. *J. educ. Psychol.*, 1940, **31**, 253–260.

using the sample mean as an estimate of the population mean. It is true that while the sample σ systematically (*i.e.*, in the long run) underestimates the population $\tilde{\sigma}$, the sample M is an unbiased estimate of the population $\tilde{M}$. It does not coincide with the population mean, except by chance, but it overestimates $\tilde{M}$ as often as it underestimates it.

Other Relations of σ and $\tilde{\sigma}$.—While formula (9.2) is the practical one to use for estimating the standard error of the mean, there are other relationships that may prove to add meaning to this discussion if they do not contribute formulas that are of practical utility from time to time.

If we have already computed the sample σ and have lost the data on sums of squares, we may still estimate $\tilde{\sigma}$ if we care to do so from the formula

$$\tilde{\sigma}_{\text{est}} = \sigma\sqrt{\frac{N}{N-1}} \qquad \text{(Population } \tilde{\sigma} \text{ estimated from sample } \sigma) \qquad (9.4)$$

where the symbols are as defined previously.

If we divide formula (9.4) through by σ, we have

$$\frac{\tilde{\sigma}_{\text{est}}}{\sigma} = \sqrt{\frac{N}{N-1}} \qquad \text{(Ratio of population } \tilde{\sigma} \text{ to sample } \sigma) \qquad (9.5)$$

And if we square both sides of this equation, we have

$$\frac{\tilde{\sigma}^2_{\text{est}}}{\sigma^2} = \frac{N}{N-1} \qquad \text{(Ratio of population variance to sample variance)} \qquad (9.6)$$

In other words, the ratio of the population variance to the sample variance is the ratio of N to $(N - 1)$. The larger N becomes, the more closely $(N - 1)$ approaches N and consequently the more closely σ approaches $\tilde{\sigma}$.

If we are not interested in knowing the size of $\tilde{\sigma}$ or even of σ, we may estimate $\tilde{\sigma}_M$ directly from the sum of squares by the formula

$$\sigma_M = \sqrt{\frac{\Sigma x^2}{N(N-1)}} \qquad \text{(Standard error of a mean directly from sum of squares)} \qquad (9.7)$$

in which the symbols are as previously defined.

Interpretation of the Standard Error of a Mean.—We are now ready to apply the standard-error formula to a concrete instance. To revive an old illustration, the ink-blot data, we find that σ is 10.45, and N is 50. Applying formula (9.2), $\sigma_M = 10.45/\sqrt{49} = 10.45/7 = 1.49$. The standard error of the mean of the ink-blot scores is 1.49, or 1.5. What we are asking when we compute this standard error is how far from the population mean the sample means like the one we obtained would vary. We

do not know what the population mean is, but from the value 1.5, we conclude that means of samples of 50 cases each would not deviate from it in either direction more than 1.5 units about two-thirds of the time. The interpretation of a standard error of a mean is in the latter respect like that of a standard deviation of a sample. The range from -1σ to $+1\sigma$ in both cases includes about two-thirds of the cases when the distribution is normal. Here we see the definite advantage of being able to assume a normal form of distribution for the means. In the sample, we know the value of the mean about which the cases vary, however, whereas in the distribution of means we do not know the value of the population mean about which those means vary.

What is the good, then, of knowing the standard error of a mean? There are several answers to this question. If we know that two-thirds of *all* sample means probably do not deviate more than 1.5 units from the population mean, we know that our *obtained* mean is probably not more than 1.5 units distant from the population mean. Since *two*-thirds of such sample means are probably not over 1.5 units from the population mean and *one*-third of them are probably more than 1.5 units from it, we can also say that the odds are 2 to 1 that the obtained mean does not differ from the population mean by more than 1.5 units. We have thus bracketed the estimate of the population mean to this extent. The smaller the σ_M, the narrower is this bracketing and the greater confidence we have in an obtained mean as an estimate of the population mean. Thus our degree of confidence in a statistic is related clearly to the size of its *SE*. The larger the *SE*, the less confidence we have in the statistic. The statistic still describes the *particular sample*, however, even when the *SE* is relatively large. But our confidence in generalizing from it depends upon its *SE*.

The odds of 2 to 1 are not regarded as heavy odds in statistics, though they may be so considered in gambling. This standard of confidence is usually regarded by statisticians as being entirely too low. We ordinarily want a great deal more assurance concerning a statistical result. If we allow wider margins, let us say of 2σ either way in the normal distribution, we have approximately 95 per cent of the sample means included and approximately 5 per cent (2.5 per cent in each tail of the curve) beyond those limits.[1] In the ink-blot data the deviations at $\pm 2\sigma$ are 2.98 units, or approximately 3.0 units, distant from the population mean.

[1] Strictly speaking, the z distance from the mean that includes the middle 95 per cent of the cases in a normal distribution is $\pm 1.96\sigma$. In this interpretation of sampling statistics we can afford to be so rough as to ignore .04 of a σ unless some very refined decision is at stake.

We could say that there are only 5 chances in a hundred that a sample mean (when N is 50) will deviate more than 3 units (in *either* direction) from the population mean. We could also say that the odds are about 19 to 1 that a sample mean will not be so far as 3 units distant from the population mean.

Let us apply the interpretation of σ_M to some other data. The practical usefulness of a statistic is often more apparent when comparing the same statistic derived from different data. In Table 9.1 are given means of Army General Classification Test scores for samples derived from different civilian occupational groups. For the sake of an illustration, we will assume that each occupational group represents a different population, as designated, and that the sampling of scores was random. What do the standard errors in this table tell us?

TABLE 9.1.—COMPARISON OF MEANS OF SCORES ON THE ARMY GENERAL CLASSIFICATION TEST AS APPLIED TO MEN FROM DIFFERENT CIVILIAN OCCUPATIONAL CATEGORIES*

Occupation	N	M	σ	σ_M
Accountant	172	128.1	11.7	0.88
Lawyer	94	127.6	10.9	1.13
Reporter	45	124.5	11.7	1.76
Sales clerk	492	109.2	16.3	0.74
Plumber	128	102.7	16.0	1.42
Truck driver	817	96.2	19.7	0.69
Farm hand	817	91.4	20.7	0.72
Teamster	77	87.7	19.6	2.23

* From Harrell, T. W., and Harrell, M. E. Army General Classification Test scores for civilian occupations, *Educ. & Psychol. Meas.*, 1945, **5**, 229–240. By permission of the publisher.

The mean in which we would have the greatest confidence, as representing the status of the general occupational population, is that for the truck driver. The odds are about 2 to 1 that this sample mean of 96.2 does not deviate more than .7 from the mean of all truck drivers that this sample represents. We could be practically certain (allowing a margin of $\pm 3\sigma$) that the obtained mean for truck drivers is not over 2 units distant from that of all truck drivers of this kind. The mean in which we have least confidence is that for teamsters by reason of its σ_M of 2.23.

Incidentally, the relation of σ_M to both σ and N can be seen roughly by comparison of the data for the occupational groups. On the whole, the largest standard errors come for samples where N is smallest—for lawyer, reporter, plumber, and teamster—though the rank orders are not perfect within this list of four. Where sample sizes are comparable, as for lawyer and teamster, and for accountant and plumber, the value for σ_M is more

apparently in proportion to the standard deviation of the sample. It can be seen that if the sample is large enough, the margin of error in an obtained mean (the margin being measured by the standard error) can be brought below one scale unit.

The Accuracy of Published Means.—The last paragraphs illustrate the point that any obtained mean really stands for a *region* rather than for a point, when it is used to estimate a population mean. This suggests that when a mean is so used we pay some attention to its standard error when deciding how many decimal places to report. If the standard error is greater than one scale unit, is there much use in reporting the mean to one or two decimal places? In this connection, the author recalls the mathematician who chided a graduate student for reporting coefficients of correlation to four decimal places "when the last three digits are probably wrong."

Statisticians are not agreed upon the rules governing the number of places to report when the standard error is considered. Kelley has proposed the rule that a published statistic should be terminated with "the decimal place given by the first figure of one-third of its standard error."[1] By "first figure," Kelley evidently means "first *significant* (nonzero) figure." Let us apply this rule to the means in Table 9.1. One-third of the σ_M for accountants is .29. The first significant figure is in the first decimal place, hence the mean may be reported to one decimal place. Even for the teamsters, where one-third of σ_M is .74, the first digit does not go into the unit column and hence we may report the mean to one decimal place.

As Kelley points out, there are 74 chances in 100 that a deviation as large as one-third of a standard error can occur by random sampling. Thus, errors greater than this standard are more likely than not to occur. It is the author's view that one might better require a limit of *two*-thirds of σ_M, beyond which about 50 per cent of the sample means would be expected. With this standard, the mean for teamsters in Table 9.1 would be reported to the nearest whole number; 88 rather than 87.7.

It should be remembered that this discussion applies only to the use of sample means used to describe populations. The mean used to describe any particular sample might well be reported to one digit beyond the last in the observed measurements, as was recommended in Ch. 4. Even when taken to represent population values, means may justifiably be reported to more places than the rules just mentioned would permit if it is believed that some reader may want to use those values for checking or for fur-

[1] Kelley, T. L., Fundamentals of statistics. Cambridge: Harvard University Press, 1947. P. 223.

ther computations. It is *most* important to keep these rules in mind when we interpret statistics that we read, when they are intended to indicate population values.

Means of Future Samples.—Note that the interpretations of σ_m given above say nothing about the means of future samples drawn from the same population and how far they may deviate from the *obtained* mean. For all we know, the one we obtained may be the highest or the lowest within the total range of obtainable sample means. *The dispersion of sample means is always around the population mean as the point of reference;* never, except by rare chance, around the obtained mean. We cannot make any very accurate prediction about where future sample means will fall, therefore, though, of course, they will be expected somewhere near the obtained mean. If we knew the value of the population mean we could certainly make such a prediction and the standard error would inform us of the probable size of error of our prediction.

When Sampling Is Not Random.—It has been repeatedly stressed that sampling statistics, including standard errors, apply only when sampling has been random. The reason for this is that the mathematics of the situation are exact only when sampling has been random. Any condition that tends to interfere with randomness of selection of observations, therefore, will make the estimation of standard errors and their application in drawing conclusions inaccurate, if not misleading. There are several noteworthy situations that depart from the random requirement. Some would lead to standard errors that are too small to describe the actual distributions of means, and others would lead to standard errors that are too large. In the former error, we would have too much confidence in the accuracy of the mean, and in the latter case we would have too little. There have been developed certain variations in the standard-error formulas to take care of some of the special situations.

Samples with Bias.—The effect of biased sampling upon the distribution of means can be strikingly illustrated by reference to some data on the training of pilots in the AAF during World War II. All pilot students were given a battery of classification tests from which was derived for each man a "pilot stanine" or composite pilot-aptitude score. Every month at the completion of preflight training, students were formed into class groups, each sent to a different primary flying school. In one study which covered a six-month period, 269 such classes had been sent to 58 training schools divided among three AAF Flying Training Commands. The mean stanine for approximately 52,000 students was 5.56. This value may be taken as the population mean in this situation. The standard deviation of the population was assumed to be 1.96. The

average size of sample (each class group in a single school) was 195.[1] From this information, using formula (9.1) we compute a standard error of 0.14. From this we would expect two-thirds of the 269 mean stanines to deviate not more than 0.14 from 5.56, if the sampling had been random. What are the facts?

When the 269 means were actually compiled in a frequency distribution and their standard deviation computed, the dispersion of means was actually found to be very much larger than was expected (see Table 9.2).

TABLE 9.2.—SAMPLING STATISTICS CONCERNING 269 CLASS GROUPS OF PILOTS IN PRIMARY TRAINING DURING A PERIOD OF SIX MONTHS IN THREE TRAINING COMMANDS OF THE AAF DURING WORLD WAR II*

Variable	Expected results		Obtained results		
	σ_M	Range	M	σ	Range
Pilot stanine	0.14	5.2–6.0	5.56	0.37	4.6–6.9
Graduation rate	3.4	56–75	65.3	9.5	40–90
Validity coefficient	.073	0.32–0.74	0.53	.088	0.21–0.71

* Including the pilot stanine, or composite pilot-aptitude score; the graduation rate or percentage of a class graduating; and validity coefficient, a biserial coefficient of correlation between stanine and graduation versus elimination.

Where one would expect a range of means within the limits 5.2 to 6.0, the actual range was from 4.6 to 6.9. Where the expected standard deviation of the distribution of means was 0.14, the actual standard deviation was 0.37. A comparison of the expected and obtained distribution of means is shown in Fig. 9.3.

The obvious conclusion is that the sampling of aviation students in pilot classes was most probably not random. One can surmise some of the causes after looking into the procedures by which class groups were made up. In each preflight class (*i.e.*, each month) a small percentage of students would fail to pass the curriculum successfully and would be held over probably to qualify for flight training in the next class. There was a tendency for the "holdovers" to be sent together to the same flight schools. They tended to be of low pilot aptitude. There may have been some geographical differences in pilot aptitude which would tend to make the averages of stanines differ systematically somewhat from one Command to another. This hypothesis could be subjected to experimental check by comparing Command averages. There were probably other reasons

[1] Actually, some classes deviated from 195 in number. For the sake of an illustration, however, we may treat the samples as if they were of constant size.

for students of similar aptitudes to gravitate together, hence the biasing of samples.

Another study was made of the graduation rates (percentage of a class group graduating) in different samples. The pertinent data are given in Table 9.2. From the over-all graduation rate of 65.3 and the size of sample, we would expect (by formula 9.18) a standard deviation of the distribution of the 269 rates to be 3.4. Actually it was 9.5. Since the probability of graduation for any cadet was strongly correlated with his aptitude score, we would expect the bias in sampling on aptitude to be reflected in biased samples as to graduation rate. This is probably not the whole story, however. There were many other conditions which could contribute to marked variations in graduation rate besides the variations in aptitude. Weather conditions varied from school to school and from month to month. Training practices and policies may have varied, in spite of close regulation. Instructor and test-pilot judgments were not standardized hurdles and may have varied from school to school.

A third study is mentioned now for comparison, although it involves the sampling errors of coefficients of correlation which are treated later. This study is concerned with the variation in validity coefficients in the same 269 class groups. The validity of the pilot stanine for predicting the training success of pilots was indicated by what is known as the biserial coefficient of correlation (see Ch. 13). This has the same value as a Pearson product-moment r, but is computed when one of the variables, assumed to be normally distributed actually, is forced into two categories. The two categories for the training criterion were the graduates and the eliminees. The standard error for a biserial correlation equal to .53 when the size of sample is 195 amounts to .073 (computed by formula 13.8). The expected and obtained statistics are given in Table 9.2 and illustrated in Fig. 9.3. In drawing the distribution curve, normal distribution of the coefficients was assumed, whereas the expected distribution should be slightly negatively skewed. The obtained distribution of the 269 coefficients was actually so skewed. At any rate, since the obtained standard deviation was only .088 and not so very different from the expected one (.073), we may conclude that if there was biased sampling with respect to the validity of pilot stanines it was of minor importance. This is reassuring for the stability of useful selection by means of the aptitude score. While there were seemingly enormous variations in validity from school to school and from time to time, amounting to a spread from .21 to .71, those variations may be regarded as due mostly to sampling errors. Incidentally, this example shows just how much obtained correlation coefficients may deviate from the population param-

eter even with samples as large as 195. Any single obtained coefficient may be anywhere in the range of such a distribution, but the saving feature is that extreme deviations are highly improbable and small ones most probable. These illustrations should demonstrate more clearly some

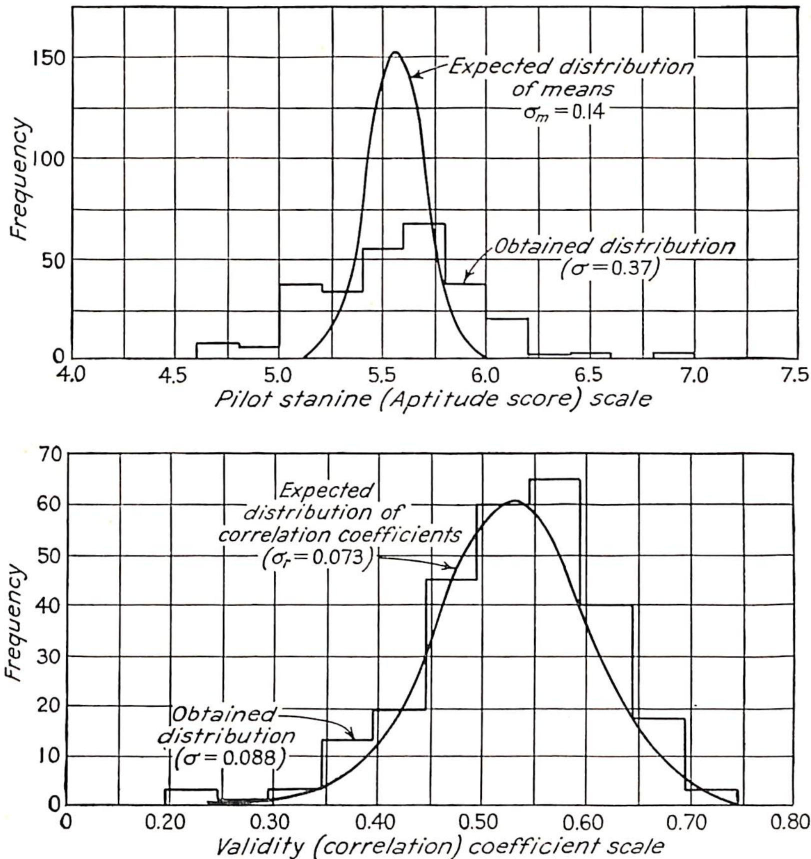


FIG. 9.3.—Distribution of expected and obtained sample means, also of expected and obtained validity coefficients, in connection with 269 samples (class groups) of AAF pilots in primary training during a five-month period in about 60 different schools. Especially to be noted is that the obtained distribution of means was much wider than expected, indicating nonrandom sampling, while the distribution of validity coefficients was about as expected, indicating random sampling. This is possible because two different kinds of sampling are involved.

of the practical uses of standard errors, as well as the importance of random sampling if we are going to make accurate and useful interpretations.

When Observations Are Not Independent.—Random sampling also implies independence of observations. In the preceding examples, observations were not independent because certain restricting conditions tied cases

together; if one student was chosen to go to a certain school at a certain time, one or more others like him were also chosen with him. There are other situations where this occurs, many times without the investigator's being aware of it. It is most likely to occur when sampling is obtained from subgroups of the population.

Suppose we have an experiment in which there are 10 subjects and each has 10 trials in each experimental session. For each session we do not have 100 independent observations. Nor do we have merely 10 observations. Because there are individual differences, the 10 observations in each set will be somewhat homogeneous, having been derived from a single source. In the larger setting of the 100 observations, they are not independent. In computing a σ_M for the mean of these 100 observations, the number of degrees of freedom is not 99. It is difficult to say just what it should be. The most conservative approach would be to assume 10 observations, each being the mean derived from one individual, and 9 degrees of freedom. But this would lead to an overestimate of the standard error. In the situation described, we have what is called *cluster sampling*. For a special treatment of this subject which includes formulas for estimating σ_M, the reader is referred to a discussion by Marks.[1]

When Populations Have Been Stratified.—Another instance of sampling from subgroups of a population is that of stratified-random sampling. The effects of grouping in this instance are in the opposite direction of those mentioned above. Stratifying tends to stabilize the dispersion of sample means and of other statistics, preventing their scattering as much as would be true of a completely random sample. Consequently, the σ_M derived in the ordinary manner is an overestimate. Such a standard error is therefore a conservative estimate of statistic fluctuations.

Certain corrective procedures have been developed for the case of stratified-random sampling. The most general and serviceable formula is

$$\sigma_M = \sqrt{\frac{\sigma^2 - \sigma^2_m}{N - 1}} \quad \text{(Standard error of mean corrected for stratification in sampling)} \qquad (9.8)$$

where σ^2 = variance in the total sample.

σ^2_m = variance among the means of the subgroups.

Each subgroup is a sample representing a stratum, within which there has been random sampling. It should be pointed out that the variance σ^2_m is a weighted affair, that is, the contribution of each set of data to the variance is in proportion to its size. The formula for this is

[1] Marks, E. S. Sampling in the revision of the Stanford-Binet Scale. *Psychol. Bull.*, 1947, **44**, 413–434.

$$\sigma^2_m = \frac{1}{N} [N_1(M_1 - M)^2 + N_2(M_2 - M)^2 + \cdots + N_k(M_k - M)^2] \quad \text{(Weighted variance of means of sample sets)} \quad (9.9)$$

where $N_1, N_2, \ldots, N_k$ = numbers of cases in sets 1 to k, respectively.
N and M refer to the total, composite sample.

For further discussion of this topic, and additional formulas, the reader is referred to an excellent treatment of the general subject of sampling by McNemar.[1]

The Size of Population.—In previous paragraphs we have assumed that populations are of infinite size; at least that they are extremely large as compared with the size of sample extracted. In some situations the total population may be finite, and not many times larger than the sample. This restriction means that successive samples would have in them many more cases in common, and this leads to greater similarity in means. If the size of the population is known, we can take it into account in estimating $\tilde{\sigma}_M$ and hence obtain a more realistic figure for it. A serviceable formula is

$$\sigma_M = \frac{\sigma}{\sqrt{N-1}} \sqrt{1 - \frac{N}{N_P}} \quad \text{(Standard error of mean corrected for size of population)} \quad (9.10)$$

in which N_P is the number in the total population and other symbols are as previously defined. It can be seen that as N_P becomes very large compared to N the correction factor under the radical at the right approaches 1.0 and the standard error of a mean is then estimated by the customary formula. When the sample contains $\frac{1}{100}$ of the population, the value of the factor at the right reduces to .995 and the standard error is only one-half of one per cent lower than it would be without the correction.

Matching of Samples.—In some investigations, there is restriction in sampling brought about by matching. Experimental and control groups are often equated in some respects while studying the effect of some varied condition upon a measured outcome. Groups are frequently "equated" for such matching variables as chronological age, mental age, *IQ*, socio-economic status, or for initial score on some particular task or test. As in the case of stratified sampling, it pays to match samples only on variables that are correlated with the measured variable—the variable on which we note the experimental outcome. The matching may be by pairs (*e.g.*, for every individual of a certain kind in the experimental group there is a similar one in the control group) or by total group (assuring that the

[1] McNemar, Q. Sampling in psychological research. *Psychol. Bull.*, 1940, **37**, 331–365.

means, standard deviations, and skewness are practically the same for the matching variable in the two groups). It is logical that if we try to keep successive samples constant with respect to the mean on some variable positively correlated with the experimental variable, the means on the latter will also be kept more constant depending upon the extent of that correlation. The standard error of a mean should then be smaller under this restriction. The general formula is

$$\sigma_M = \frac{\sigma}{\sqrt{N-1}}\sqrt{1 - r^2_{mx}} \tag{9.11a}$$

(Standard error of a mean corrected for effects of matching samples)

$$= \sigma\sqrt{\frac{1 - r^2_{mx}}{N-1}} \tag{9.11b}$$

where r_{mx} = correlation between the matching variable and the experimental variable.

Inspection of formula (9.11*a*) will show that the first factor, $\sigma/\sqrt{N-1}$, is the customary standard error. What the second factor, $\sqrt{1 - r^2_{mx}}$, does is to modify, by lowering, the size of the standard error. The larger r becomes, the greater is the correction effect. The correlation has to be as high as .866 in order to make the correction as much as .50, in which case the standard error is half as large as it would be without matching. The same change in σ_M could be accomplished by increasing the size of sample four times with *random* sampling. When r_{mx} is .707, the reduction is equivalent to that obtainable by doubling the size of sample in random sampling. This gives some idea of the economy of measurement to be achieved by matching samples.

If the matching has been done on the basis of more than one variable, the correlation called for in formula (9.11) is the multiple correlation (see Ch. 16) between a combination of the matching variables and the experimental variable. Matching on the basis of many variables does not ordinarily pay unless the matching variables are themselves relatively independent, *i.e.*, uncorrelated with each other. Adding one more matching variable to several may not increase the multiple correlation and hence not lower the standard error.

Sometimes a sample group is matched on the *same* variable, as when we give it a pre-test and a post-test, with intervening experience or practice. In this case, the paired cases are identical individuals. The variability of means to be expected from successive sampling of this kind is indicated by the following estimate of the standard error:

$$\sigma_M = \frac{\sigma}{\sqrt{N-1}}\sqrt{1 - r_{xx}} \tag{9.12}$$

(Standard error of a mean for matching a group on the experimental variable)

in which r_{xx} = the test-retest reliability (see Ch. 17) of the experimental variable. The reader who is familiar with the reliability statistics described in Ch. 17 will recognize the product $\sigma \sqrt{1 - r_{xx}}$ as the *standard error of measurement* of individuals. Dividing by degrees of freedom should indicate similarly the dispersion of means of measurement.[1]

The Reliability of a Median.—The variability of sample medians is about 25 per cent greater than the variability of means when the population is normally distributed. Under this condition the standard error of a median can be estimated by the formula

$$\sigma_{Mdn} = \frac{1.253\sigma}{\sqrt{N}} \quad \text{(Standard error of a median estimated from } \sigma) \qquad (9.13)$$

in which σ_{Mdn} stands for the standard error of a median. As applied to the ink-blot test data,

$$\sigma_{Mdn} = \frac{(1.253)(10.45)}{\sqrt{50}} = \frac{13.09385}{7.071} = 1.85$$

Two-thirds of the sample medians of ink-blot scores, when N equals 50, in samples drawn at random from the population will be expected within 1.85 units of the population median. Since the population is normally distributed, by assumption, we may also say that the sample medians would not deviate from the population *mean* more than 1.85 units, two-thirds of the time. The median may thus be used as an estimate of the population mean, but with less confidence than we have in the use of the sample mean for the same purpose.

THE RELIABILITY OF OTHER STATISTICS

The Standard Error of a Standard Deviation.—The standard deviation will also fluctuate from sample to sample. For a given size of sample, the sampling distribution of σ is somewhat skewed for small samples but approaches the normal form so closely for large samples that we can draw inferences about a sample σ, knowing its standard error. This *SE* is estimated by the formula

$$\sigma_\sigma = \frac{\sigma}{\sqrt{2N}} \quad \text{(Standard error of a standard deviation)} \qquad (9.14)$$

Applied to the ink-blot data,

$$\sigma_\sigma = \frac{10.45}{\sqrt{100}} = 1.045$$

[1] For further discussion of standard errors in matched and other restricted samples see Peters, C. C., and Van Voorhis, W. R. Statistical procedures and their mathematical bases. New York: McGraw-Hill, 1940. Pp. 132–135.

We can now say that the odds are 2 to 1 that the sample σ will not deviate more than 1 unit (1.045 should be rounded to 1.0) from the population $\tilde{\sigma}$. We can also say that the odds are about 19 to 1 that a sample σ will not deviate more than 2 units from $\tilde{\sigma}$.

Comparing formula (9.14) with formula (9.2) for the standard error of a mean, we can see that a population standard deviation is more accurately estimated than a population mean, when we compare them as to sampling errors. The denominators of these two formulas contain the values $2N$ and $(N - 1)$, respectively, which means that the σ_M is usually more than 40 per cent greater than σ_σ. For the inkblot data, the two standard errors are 1.045 and 1.49, respectively. In one sense it is fortunate that the standard deviation is more stable than the mean, because both σ_M and σ_σ are estimated from it.

When there are departures from ordinary random sampling—from stratified populations, finite populations, or from matched samples—corrections in the estimate of σ_σ are in order just as they are in connection with σ_M. By analogy to formulas given above for σ_M, one can make the appropriate corrections in σ_σ. In general, the occasion for computing a σ_σ is a rare event. The need for making a correction in one of these special situations is even more rare. But where called for, as in the case of σ_M, such a correction may make a real difference in conclusions drawn.

The Standard Error of Q.—The standard error of the semi-interquartile range is estimated by the formula

$$\sigma_Q = \frac{.7867\sigma}{\sqrt{N}} \qquad \text{(Standard error of } Q \text{ estimated from } \sigma) \qquad (9.15)$$

when the population distribution is normal. Applied to the ink-blot data,

$$\sigma_Q = \frac{(.7867)(10.45)}{\sqrt{50}} = \frac{8.221}{7.071} = 1.16$$

If the standard deviation is not known, the next best procedure is to use the formula

$$\sigma_Q = \frac{1.166Q}{\sqrt{N}} \qquad \text{(Standard error of } Q \text{ estimated from } Q) \qquad (9.16)$$

This substitute formula is possible because in a normal distribution $Q = .6745\sigma$. Applied to the ink-blot data, this formula gives

$$\sigma_Q = \frac{(1.166)(7.5)}{\sqrt{50}} = 1.24$$

The slight discrepancy between σ_Q as estimated by these two formulas may be due to the fact that the sample distribution was not quite normal and hence Q did not equal exactly $.6745\sigma$, or to minor irregularities in frequencies in class intervals that were crucial for the estimation of Q.

The interpretation of σ_Q is comparable to that of other standard errors already encountered in this chapter, *i.e.*, in terms of degree of confidence that the sample Q could deviate certain distances from the central value of the sampling distribution.

The Reliability of a Proportion.—Data in terms of frequencies, percentages, and proportions are so common in psychology and the social sciences that the problem of their reliability is very important. Each obtained proportion is a sample statistic and, as such, it may be expected to fluctuate from sample to sample. Out of 100 students quizzed at random, the proportion of them who reported the habit of reading a daily newspaper is .65. How well does this proportion represent the student population? Assuming that we have a random sample, there is a way of estimating how such a proportion of 100 observations might be expected to vary. The standard error of a proportion measures this variation, and with a known or assumed form of distribution of the sample proportions we can arrive at conclusions as to the accuracy of the obtained result.

The standard error of a proportion is given by the formula

$$\tilde{\sigma}_p = \sqrt{\frac{\tilde{p}\tilde{q}}{N}} \qquad \text{(Computed standard error of a proportion)} \qquad (9.17a)$$

where $\tilde{p}$ = the proportion of the *population* who are in the category selected.

$\tilde{q}$ = the proportion of the *population* who are not in the category ($\tilde{q} = 1 - \tilde{p}$).

N = the number in the *sample*.

We ordinarily do not know the parameters $\tilde{p}$ and $\tilde{q}$. The practical solution is to use the sample p and q as the best estimates we know for those values.

The useful formula is therefore

$$\sigma_p = \sqrt{\frac{pq}{N}} \qquad \text{(Estimated standard error of a proportion)} \qquad (9.17b)$$

The total outcome of formula (9.17) depends relatively more upon the size of N than upon $\tilde{p}$ and $\tilde{q}$ because the product pq remains fairly constant between .20 and .25 for quite a range of values of p (namely, between .27 and .73) and because in most cases the sample p will not be very divergent from $\tilde{p}$. As p goes outside the limits of .27 to .73 and

as it approaches 0.0 or 1.0, the divergence of p from $\tilde{p}$ becomes smaller and smaller. If one has better knowledge concerning the population $\tilde{p}$, which is provided by other information, for example a p from a larger sample or from a series of prior samples, one could use some other estimate of $\tilde{p}$ as a hypothesis. One could arbitrarily choose some hypothetical $\tilde{p}$ derived upon the basis of a priori reasoning. This approach will be given more attention in Ch. 11 on "Testing Hypotheses," so will not be discussed further here.

For the newspaper-reading data suggested above, where p is .65 and N is 100, the standard error is therefore estimated by formula

$$\sigma_p = \sqrt{\frac{(.65)(.35)}{100}} = \sqrt{.002275} = .048$$

The interpretation of this result, as usual, depends upon an assumption about the form of the sampling distribution. The sampling distribution of p approaches the normal form if N is not too small, and if p is not too close to .00 or to 1.00. It must be stated by way of qualification that as $\tilde{p}$ deviates from .50 in either direction the distribution of p becomes skewed. This is because no p can fall below 0.0 or go above 1.0. Distributions are curtailed at those extremes but can extend greater distances in the opposite direction. As samples become very large, however, dispersions become so narrow that these terminal restrictions have less importance. As a practical rule for avoiding seriously nonnormal sampling distributions of p, some statisticians recommend that we forgo estimating σ_p, or at least interpreting it, when the product Np (or Nq, whichever is smaller) is less than 10.[1] Thus, if N is as small as 20, only one proportion could qualify to meet this rule, namely $p = .5$. For small samples greater than 20 there is less restriction, but some. For example, if $N = 40$, only proportions between .25 and .75 could qualify for meeting normal-distribution standards under this rule. There are other methods of dealing with cases that do not come under this rule.[2]

The obtained σ_p in connection with the newspaper data is .048, or approximately .05. Since the conditions for normal distribution of the sample proportions are apparently satisfied, we can say that the odds are about 2 to 1 that the obtained proportion is not further than .05 from the population proportion. Our margin of error in the proportion of .65 may be stated as .05. Enlarging our confidence limits, we may feel much

[1] Treloar, A. E. Elements of statistical reasoning. New York: Wiley, 1939. P. 180.

[2] *Ibid.*, Ch. 12.

more certain (odds about 19 to 1) that this obtained proportion is not more than .10 away from the population value.

The Proportion as a Mean.—In connection with the question of reliability of a proportion it is interesting to know that in one important sense the proportion is actually a mean and its standard error is actually the standard error of a mean. A numerical example will illustrate this point.

Suppose we have administered a certain test item to 100 individuals, of whom 80 give the correct answer and 20 do not. Let each successful person receive a "score" of 1 and each unsuccessful person a "score" of 0. That is actually what we usually do in scoring a test composed of items. Each item may be regarded as a subtest on which the range of scores is usually 2 units. We need not confine this reasoning to responses to test items. Wherever events can be classified into a certain category or not, we can arbitrarily give a value of 1 to all cases in the category and a value of 0 to those not in the category. Other examples might be possessing a habit of reading a daily newspaper versus not having the habit; being an alcoholic versus not being an alcoholic; voting for candidate X versus not voting for candidate X; and so on. In terms of probability, the value of 1.0 stands for absolute certainty of an event's occurring and zero stands for absolute certainty of its not occurring. A proportion can thus be regarded as an *average probability*.

Returning to the test-item problem, the mean score for the 100 individuals is the sum of the scores divided by the number of them, in other words, $\Sigma X/N$ or $\Sigma fX/N$. The sum of the scores is 80 and N is 100, from which the mean is .8. This is also the proportion passing the item. Thus our proposition that the proportion is a mean is demonstrated.

To find the standard error of a mean, as by formula (9.2), we need to know the standard deviation of the sample. It can be shown that for a distribution in two categories the variance is equal to the product pq and the standard deviation is equal to $\sqrt{pq}$. This is demonstrated in Table 9.3. This table shows both the numerical solution for this particular illustrative problem and also the general solution in terms of symbols. From the table it should be clear that the variance equals pq and the standard deviation equals $\sqrt{pq}$. Using the latter as an unbiased estimate of the population standard deviation, by substitution for σ in formula (9.2) we have $\sqrt{pq}/\sqrt{N}$, or $\sqrt{pq/N}$, which is formula (9.17*b*) for the standard error of a proportion. Having used the obtained p as an unbiased estimate of the population $\tilde{p}$, since it is a mean, we need not be concerned with loss of degrees of freedom here. Consequently the denominator in formula (9.17) is $\sqrt{N}$ rather than $\sqrt{N - 1}$.

TABLE 9.3.—COMPUTATION OF THE MEAN AND STANDARD DEVIATION FOR A DISTRIBUTION IN TWO CATEGORIES

	Numerical example					Solution with symbols			
	X	f	fX	x	fx^2	f	fX	x	fx^2
	1	80	80	+0.2	3.20	Np	Np	q	Npq^2
	0	20	0	−0.8	12.80	Nq	0	$-p$	Np^2q
Sum		100	80		16.00	$Np + Nq = N(p + q) = N$	Np	—	$Npq^2 + Np^2q = Npq(p + q) = Npq$
Mean			.80 (M)		.16 (σ^2)		p (M)		pq (σ^2)
Standard deviation					.4				$\sqrt{pq}$

The Standard Error of a Percentage.—If we wish to work in terms of percentages instead of proportions we may do so. Let the percentage be denoted by P and let Q equal $100 - P$. Remembering that a percentage is 100 times its corresponding proportion, the standard error of a percentage will be 100 times as large as that for the proportion. The formula reads

$$\sigma_P = 100\sqrt{\frac{pq}{N}} = \sqrt{\frac{PQ}{N}} \quad \text{(Standard error of a percentage)} \quad (9.18)$$

The Standard Error of a Frequency.—A frequency, or the number of cases in a certain category, is equal to N times p, the proportion; consequently the standard error of a frequency is N times that for a proportion, and we have the formula

$$\sigma_f = N\sqrt{\frac{pq}{N}} = \sqrt{Npq} \quad \text{(Standard error of a frequency)} \quad (9.19)$$

Out of 30 students who attempted a certain test item, 18 succeeded and 12 failed. How much confidence can we have that the 18 successes represent the actual success rate for the larger population these 30 students represent? The standard error, assuming a population $\tilde{p}$ equal to .60, by formula (9.19) is equal to $\sqrt{30 \times .6 \times .4} = \sqrt{7.20} = 2.7$. This obtained frequency may therefore be presumed not to deviate more than 2.7 from the average frequency to be expected if we had examined the entire population in samples of 30, with a degree of confidence that can be expressed as a 2 to 1 bet. With a degree of confidence expressed by a 19 to 1 bet, we could say that we do not expect that this obtained fre-

quency departs more than 5.4 from the average frequency we would get from many such samples.

Standard Errors of Proportions When Sampling Is Not Completely Random.—When sampling has been stratified or clustered or when populations are restricted in size, we need to make corrections analogous to those already proposed for standard errors of other means. This holds also for standard errors of percentages and of frequencies. Corrections for the latter can be obtained because of their relations to the former, as indicated above.

When there has been stratification, the standard error of a proportion is estimated from the formula

$$\sigma_p = \sqrt{\frac{pq}{N} - \frac{\sigma^2_m}{N}} \quad \text{(Standard error of a proportion corrected for stratification)} \quad (9.20)$$

where p = proportion observed in the entire sample, all strata combined.
$q = 1 - p$
N = number in the total sample.
σ^2_m = weighted variance of the several strata proportions about the total sample proportion, p.

The solution for σ^2_m needed in formula (9.20) is given by the formula

$$\sigma^2_m = \frac{1}{N}[N_1(p_1 - p)^2 + N_2(p_2 - p)^2 + \cdots + N_k(p_k - p)^2]$$

(Weighted variance of sets of sample proportions) (9.21)

where $N_1, N_2, \ldots, N_k$ = numbers of cases in the different strata, respectively, there being k strata.
$p_1, p_2, \ldots, p_k$ = proportions observed in these various strata.
N and p are as defined above.

It can be seen that if the various strata proportions all equal p, or nearly so, the variance σ^2_m is zero, or nearly so, and the standard error σ_p is the same as in purely random sampling. If the variance between strata is not zero, formula (9.20) will give a smaller σ_p than would be obtained from formula (9.17).

If the population is of finite size and not too many times as large as the sample, the following formula will provide an improved estimate of σ_p:

$$\sigma_p = \sqrt{\frac{pq}{N}\left(1 - \frac{N}{N_P}\right)} \quad \text{(Standard error of a proportion corrected for size of population)} \quad (9.22)$$

in which N is the size of sample and N_P the size of the population. This

formula is clearly analogous to formula (9.10) for the similar correction of σ_M.

When samples have been matched on the basis of some outside variable correlated with the categorical variable on which the proportion is based, by analogy to formula (9.11),

$$\sigma_p = \sqrt{\frac{pq}{N}(1 - r^2_{mx})} \quad \text{(Standard error of a proportion corrected for effects of matching)} \qquad (9.23)$$

in which r_{mx} = correlation between the matching variable X_m and the experimental variable.

The correlation would be a point-biserial or a phi coefficient (see Ch. 13).

Reliability of a Coefficient of Correlation.—Like every statistic, the coefficient of correlation is subject to errors of sampling. Let us say that in a certain population the parameter correlation, $\tilde{r}$, is equal to .30.[1] From this population we take successive samples of 50 pairs of observations each. The sample r's will fluctuate in a sampling distribution around the population value, both above it and below it. An example of this has already been reported in Table 9.2, where $\tilde{r}$ was .53. How much variability may we expect? We need a standard error of r and some knowledge of the form of sampling distribution in order to say.

Sampling Distributions of r.—The sampling distribution of correlation coefficients is not of a uniform shape. It depends both upon the size of r and the size of sample. It is already known to the reader that the limits of r are -1.0 and $+1.0$. An obtained coefficient cannot exceed those limits. Consequently, as the population r approaches those limits, the sampling distribution becomes more and more skewed; negatively skewed for positive r's and positively skewed for negative r's. Only when the population $\tilde{r}$ is approximately zero is the sampling distribution expected to be symmetrical (see Fig. 9.4). For large samples, however, one need not worry very much about skewness in practice when $\tilde{r}$ is within the limits of $-.80$ to $+.80$. The larger the sample, the narrower the dispersion of r's, and consequently the less restricting effect provided by the limits of -1.0 and $+1.0$. It is conceivable that with enormously large samples, with standard deviations of .01 or .02, even when r is .90, the sampling distribution could be regarded as symmetrical with negligible discrepancies. On the other hand, even when r is zero, if the sample is very small (under 25) it is not safe to base interpretations upon the

[1] Some authors use the symbol ρ (rho) to stand for a population correlation. Since this is also the symbol used for a rank-difference correlation it seems unwise to use it here. The use of $\tilde{r}$, while not at all common, is consistent with analogous symbols—$\tilde{M}$ and $\tilde{\sigma}$, for example.

assumption that the sampling distribution is normal, for reasons which will be left to the discussion of small-sample statistics.

An Estimate of σ_r.—We can estimate the standard error of r by the general formula

$$\sigma_r = \frac{1 - r^2}{\sqrt{N - 1}} \quad \text{(Standard error of a product-moment coefficient of correlation)} \tag{9.24}$$

This formula would be more accurate if we wrote $\tilde{r}$ instead of r. There is little risk in using r as an estimate of the population parameter that is

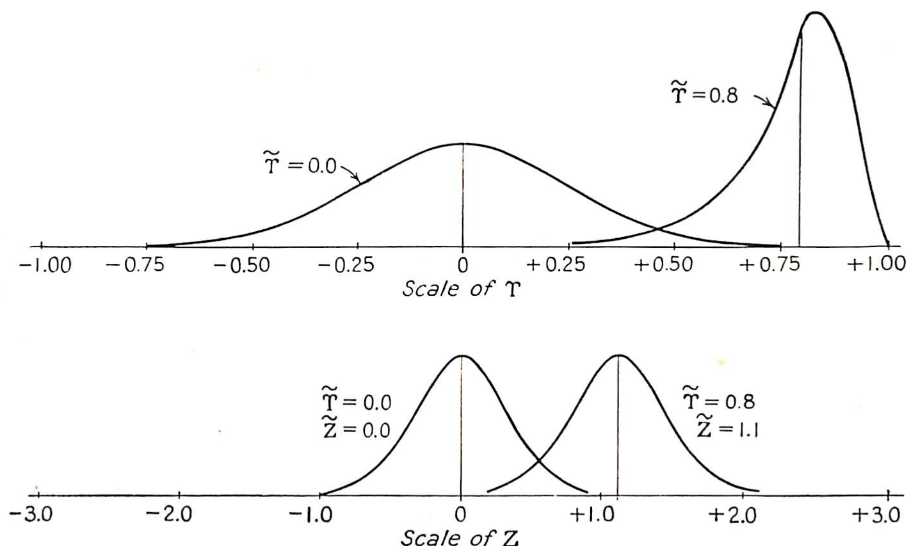


FIG. 9.4.—Distributions of sample coefficients of correlation when N is very small and when the population correlations are .00 and .80. Corresponding to them are distributions of Fisher's **z** coefficients. Conversion of r to **z** brings about symmetrical sampling distributions, regardless of the size of r.

really needed if samples are large and if r is large. Examination of the formula will show that for the same size of sample, σ_r is largest when $r = .00$ and becomes smaller as r approaches -1.0 or $+1.0$. The size of the standard error itself indicates to some extent the amount of risk we take in letting r stand for the population value.

To illustrate formula (9.24) with the case of a population $\tilde{r}$ first, let us take the values mentioned above—with $\tilde{r} = .30$, and $N = 50$. We have

$$\sigma_r = \frac{1 - .09}{\sqrt{49}} = \frac{.91}{7} = .13$$

Interpreted, this means that with a population $\tilde{r}$ equal to .30, we may expect two-thirds of samples r's, when $N = 50$, to lie within .13 of the

parameter $\bar{r}$, in other words, between .17 and .43. We also might expect 95 per cent of the sample r's under these conditions to be between .04 and .56, these values being $2\sigma_r$ distances from .30. There would be only one chance in 100 that sample r's could deviate as much as .335 (this being equal to $2.58\sigma_r$) from the population value. This much deviation marks off the range from $-.035$ to .635. We should not be too sure of these interpretations involving the extreme tails of the distribution, since departures of the sampling distribution from normal form would show up most at those places. But it can be seen how even negative coefficients might arise by random sampling occasionally, even when the population correlation is as large as .30. The smaller the r and the smaller the sample, the more likely are these reversals of algebraic sign of correlation to occur.

Consider next the case when we must substitute an obtained r for the parameter $\bar{r}$ in the use of formula (9.24). Let us use the obtained correlation of $+.61$ from the problem in Table 8.5.

$$\sigma_r = \frac{1 - .61^2}{\sqrt{87 - 1}} = \frac{1 - .3721}{\sqrt{86}} = \frac{.6279}{9.2736} = .068$$

It is sufficient to report σ_r, as for most standard errors, to two significant digits. From the result we may say that whatever the population $\bar{r}$ may be (and it is probably not far from .61), an obtained r such as .61 would not deviate from it by more than .068 with a confidence indicated by odds of 2 to 1. There are less than 5 chances in 100 that in samples of this size the sample r would depart more than .136 from the population value, and less than 1 chance in 100 that the sample r would depart more than .175, above or below it. The obtained r, consequently, seems securely placed in a region that is removed from zero or negative correlations.

The Significance of Small r's.—When r's are small, *i.e.*, in the region of zero but either positive or negative, our interest should usually center on the question as to whether such values could have arisen when the population correlation is actually zero. In the previous illustrations we were more concerned with the accuracy of determination of the *amount* of correlation. Incidental to that problem we saw that some sampling distributions could come close to zero if not extend beyond it. This becomes

a very serious problem when coefficients are numerically small and samples are not large enough to fix the boundaries of sampling fluctuation definitely clear of zero.

The best approach to the small r is to assume that the population correlation is actually zero and then ask whether, with the size of sample being what it is, the obtained r could have occurred merely by random sampling. Our being able to conclude whether the obtained r represents any genuine correlation at all depends upon this kind of test. Incidentally, assuming that the population r is zero is one form, or one application, of the *null hypothesis* of which we will hear much more later on. Our working hypothesis is that there is a *null* amount of correlation. Since formula (9.24) implies the use of the population $\tilde{r}$, we may insert any value for it that we please (except ± 1.00, which would shrink σ_r to zero). Any r we chose to insert would be our hypothesis about the amount of correlation. We could then compute σ_r and test the hypothesis by seeing whether the obtained r deviates too far from $\tilde{r}$ to be reasonable. A deviation that goes outside the practical limits of the normal distribution would of course be very unreasonable. A deviation that is so large as to occur by chance only a very small proportion of the time would also be seriously questioned.

When the population $\tilde{r}$ is zero, the standard error is estimated by the formula

$$\sigma_{r_o} = \frac{1}{\sqrt{N-1}} \qquad (9.25)$$

(Standard error of r when the population $\tilde{r}$ is assumed to be zero)

This formula will apply satisfactorily when N is not less than 25, and certainly when it is not less than 50. Applying this formula to the data of Table 8.5,

$$\sigma_{r_o} = \frac{1}{\sqrt{86}} = \frac{1}{9.27} = .11$$

The obtained correlation, .61, is more than 5 times as large as this standard error. So very rarely could this much correlation occur by random sampling in a population where X and Y are actually uncorrelated, that we can reject the null hypothesis and say that almost certainly there is positive correlation. We would not ordinarily make this test of a coefficient as large as .61 unless the sample were quite small. Even if the sample were 26, in which case $\sigma_{r_o} = .20$, this obtained correlation would be at least 3 times the standard error.

The t-ratio Test of r.—The test of the null hypothesis, as was just illustrated, lies in the examination of the ratio of an obtained r to σ_{r_o}. In a normal distribution this ratio is, of course, a standard measure, z. When we are dealing with sampling distributions, however, the custom is to give this ratio a new symbol, t, and to speak of the ratio as a t ratio. In general, *t is defined as the ratio of a deviation to a standard error.* In this case we are dealing with the deviation of an obtained r from a (assumed) population $\bar{r}$. The population $\bar{r}$ is assumed to be the mean of a sampling distribution, thus the t ratio is interpretable as a standard measure in relation to the normal distribution, when samples are large. When samples are small, as we shall find later, we have other distributions to take its place.

How large a correlation, or how large a t, is needed in order to lead us to reject the null hypothesis? There is no single standard for rejection. The reason is that values of t are on a continuous scale (see Fig. 9.5) and all we can do is to note the probability of so large a t occurring by chance. The smaller that probability, the more inclined we are to doubt the null hypothesis, if not to reject it. If we reject the null hypothesis, the chief alternative we have is to believe in a population correlation different from zero. This is one of the chief virtues of a statistical test of significance. The situation is reduced to two alternatives—either the null hypothesis, in this case, or some other. There is either some correlation (r not zero) or there is not. If we reject the null hypothesis with considerable confidence, we have strong reason to accept the other. Beyond this, as we shall see, there is weakness in this statistical test; if we do not feel justified in rejecting the null hypothesis we are not thereby forced to accept it. We can find evidence from the t test favoring the rejection of a null hypothesis but if we cannot reject it the outcome is inconclusive. These points will become clearer in subsequent discussion. We will proceed one step at a time.

Confidence Levels.—The larger the t, the less likely it is that it could occur by random sampling. There is general agreement that when t is as large as 1.96 (in normal sampling distributions) we may regard t, and the deviation for which it stands, as "significant." In a normal distribution, a t that deviates more than 1.96 (in *either* direction) from the mean would occur only 5 times in 100.[1] This criterion is often referred to as

[1] The 5 per cent is equally divided between the two tails of the curve, *i.e.*, areas of .025 in each tail. Figure 9.5 shows this fact, also the fact that for the 1 per cent level of confidence we are talking about .005 of the area in each tail. The reason why both tails are included is that with a symmetrical distribution it is as likely for a t of a certain size to occur in one direction as in the other. In other words, it is not the direction but the *size* of t that matters.

"significant at the .05 level of confidence," or as "significant at the 5 per cent level." We could reject the null hypothesis with confidence that only 5 times in 100 would we be wrong in so doing. We would be wrong if we rejected the hypothesis when actually it happened to be true, *i.e.*, there *was* actually a population correlation of zero. A more confident criterion of rejection requires a t as large as 2.58, at which value there is less than 1 chance in 100 that a t as large or larger could have occurred by chance. With such a t obtained from a sample, we could reject the null hypothesis with the confidence of being wrong only once in 100 times. This is confidence at the .01 or 1 per cent level. Other levels are often mentioned by some investigators who regard the 5 per cent and 1 per cent levels insufficiently refined as criteria. The other levels, along with these two, are summarized in Table 9.4.[1]

TABLE 9.4.—CRITERIA OF SIGNIFICANCE OR CONFIDENCE LEVELS OF t IN A NORMAL DISTRIBUTION

Level of t	Level of Confidence	Rough conclusion
Below 1.65......	Below .10 or 10% level	Insignificant
1.65......	At the .10 or 10% level	Insignificant
1.96......	At the .05 or 5% level	Significant
2.33......	At the .02 or 2% level	Significant
2.58......	At the .01 or 1% level	Very significant
2.81......	At the .005 or 0.5% level	Very significant
Above 2.81......	Beyond the .005 or 0.5% level	Very significant

By reference to Table 9.4, we note that if an obtained t were equal to 1.80, the discrepancy between hypothesis and fact is said to be significant between the 10 per cent and 5 per cent levels of confidence. If a t is 2.75 we would say that it is significant between the 1 per cent and 0.5 per cent levels of confidence. A t of 1.80, or any value below 1.96, however, is ordinarily regarded as "insignificant." We would not reject the null hypothesis unless we could do so at least beyond the .05 level, though that is an arbitrary custom. A t greater than 1.96 but less than 2.58 is often reported as being "significant" and a t of 2.58 or greater as being "very significant." An investigator may choose any level of confidence he prefers. But he must defend it. It may depend upon the kind of problem being investigated and upon the seriousness of being wrong in concluding either for or against the null hypothesis. At any rate, it is best practice for the investigator to decide upon the level of significance

[1] Confidence levels are also called *fiducial limits* by R. A. Fisher and others.

he is going to require in advance of knowing any statistical results, lest he be biased by such knowledge in this decision later.

Errors in Making a Statistical Inference.—There are two chances of coming to wrong conclusions. One is setting the level of confidence required so low that there is danger of rejecting the null hypothesis when it is actually true. The statisticians call this an *error of the first kind.* The probability of making this kind of an error is as small as the probability of a t of the size of the criterion which was adopted occurring by random sampling. Thus, if t is significant at the 2 per cent level of confidence and we reject the null hypothesis, the probability that we would be wrong in doing so is .02, or 2 chances in 100. If, however, t is significant at the 10 per cent level and we reject the null hypothesis, we have one chance in 10 of being wrong. An *error of the second kind* is in accepting the null hypothesis when it is false. The danger of this error is increased if we put the criterion too high. There is no easy method of determining the probability of this kind of error. We can reduce the chances of it by lowering the level of significance required for rejection.[1]

Another logical point must be emphasized in connection with the t test of significance. Decision not to reject the null hypothesis does not necessarily prove that it is true. We have already seen that rejection of it does not necessarily prove that it is false. We only have degrees of confidence (no final proof) that we are correct in rejecting it. If an r deviates from zero so little that the t is below the 5 per cent criterion, we do not reject the idea that the population correlation *could be* zero but this is not the same thing as saying that the correlation *is* zero. It could be actually anything within the range marked off by *chance* deviations as determined by a standard error. There is even more justification for saying that the population $\bar{r}$ is the same as the obtained r, in the absence of any other information, than for saying that it is actually zero. The point is that when any investigator obtains a very small r, if he takes it to mean actual correlation, it is incumbent upon him to show the improbability of such an r arising from variables uncorrelated in the population. He can still maintain that the two variables are correlated in his sample, and he would be right, because the least we can say about an obtained r, as about any statistic, is that it describes something about a particular sample. But using a statistic, including r, to describe a population calls for supporting evidence.

The t ratio will be encountered many times again. It is particularly useful in testing the significance of differences of many kinds. Its inter-

[1] See Deemer, W. L. The power of the t test and the estimation of required sample size. *J. educ. Psychol.*, 1947, **38**, 329–342.

pretation in small samples will also receive attention later in this chapter.

Minimum Significant r's.—A more convenient and practical procedure for determining whether an obtained coefficient of correlation is significantly different from zero is provided by the Wallace-Snedecor tables (see Table D, Appendix B). In the first column of the table are given the number of degrees of freedom available for the coefficient. In each correlation problem the number of degrees of freedom is $N - 2$. The number of observations is a *pair* of values, one in X and one in Y. One degree of freedom is considered lost in the computation of each mean, the mean of X and the mean of Y. Both the products of the moments and the two standard deviations are affected by the loss of two degrees of freedom. This leads to some bias in the sample r as an estimate of the parameter $\tilde{r}$, incidentally, but it is inconsequential unless N is small.

Having located the proper number of degrees of freedom in Table D, we find in the second column two values. One is the minimum r that is significant at the 5 per cent level, and the other, in bold-face type, is the minimum r significant at the 1 per cent level. If we are satisfied with these gross criteria for rejection of the null hypothesis regarding correlation, this procedure will do. If we want greater refinement or other standards, we would use formula (9.25), or, in the case of small samples, formula (9.38). One advantage of the use of Table D for this purpose is that it takes care of small samples as well as large samples. The minimum r's listed were derived on the basis of formula (9.38) which will be discussed under small-sample statistics.

Examination of Table D shows that for samples with 1,000 degrees of freedom r must be at least .062 to be significant at the 5 per cent level. An r of .062 or larger, positive *or* negative, could arise by chance when r is zero only 5 times in 100. If we reject the idea that the population correlation is zero, we have 5 chances in 100 of being wrong. For the same size of sample, an r of .081 is required for significance at the 1 per cent level. Thus, if we obtained a correlation of .10 (either positive or negative) we could feel very confident that there is *some* relationship between X and Y and that it is in the direction indicated by its algebraic sign. Since we feel confident that there is some degree of correlation present, we might then apply the usual formula for σ_r (9.24) in order to get an idea of its probable limits.

Thus, even very low coefficients, like .10, may indicate a relationship, but it takes a very large sample to establish that fact and to determine its probable value. On the other hand, some obtained coefficients of moderate or even large size may be very uncertain indicators of any

relationship at all. Note that when N is 10 (8 degrees of freedom), the minimum r's required are .632 and .765, at the 5 per cent and 1 per cent levels, respectively. Even if our obtained r exceeded these limits when N is 10, the exact level of the *amount* of correlation would be exceedingly uncertain. Correlations derived from such small samples are practically worthless, unless they are of the order of .90 or higher.

Fisher's **z** *coefficient.*—Because of the numerous radical departures of the sampling distribution of r from normal form, and the limitations to our interpretations that result from this, R. A. Fisher has developed another statistic into which any obtained r can be converted by formula and which has a normal sampling distribution even with very small samples. This statistic has been called **z**, which we will write in bold face to distinguish it from the standard measurement z. They are definitely not the same statistic. Fig. 9.4 shows distributions of r's and of corresponding **z**'s on their respective scales.

The range of **z** is from $-\infty$ to $+\infty$, but when r reaches the value .995, **z** is still short of the value 3.0. Up to an r of .25, **z** and r have approximately the same value. Even when $r = .50$, **z** is no larger than .56. Within these limits, then, distributions of r can be regarded as normal. Above this range, when normal distribution is an important consideration, it would be well to convert r to **z**. This conversion formula is

$$\mathbf{z} = \tfrac{1}{2}[\log_e (1 + r) - \log_e (1 - r)] \qquad (9.26)$$

(Conversion of a coefficient of correlation into Fisher's **z**)

in which $\log_e$ stands for a logarithm to the base e, or refers to the use of the Naperian system of logarithms.[1] In terms of logarithms in the common system,

$$\mathbf{z} = 1.1513\ [\log_{10} (1 + r) - \log_{10} (1 - r)] \qquad (9.27)$$

(Same as formula (9.26) in terms of common logarithms)

For general practice, Table H (Appendix B) may be used for the conversion of r to **z** and of **z** to r. One would not report final results in terms of **z**, but would finally convert back to r.

The standard error of **z**, unlike that for r, is of practically uniform size for all values of **z**. It can be estimated by the formula

$$\sigma_{\mathbf{z}} = \frac{1}{\sqrt{N - 3}} \qquad (9.28)$$

(Standard error of **z**)

[1] For the benefit of the mathematically sophisticated student, **z** is the hyperbolic arc tangent of r, or $\mathbf{z} = \tanh^{-1} r$.

The interpretation of any estimate of σ_z is like that for any other standard error. It may be used to mark off confidence limits on the scale of **z** which can be referred back to corresponding r's.

The chief uses of **z** are to be found in problems of averaging coefficients of correlation (see Table 13.13) and in testing the significance of differences between r's (see p. 224) when r's are large and sample sizes are not.

THE RELIABILITY OF DIFFERENCES

Of much more practical value than the standard errors of means, proportions, and the like are the standard errors of differences between means and between proportions and the like. In experimental practice, we are perpetually comparing measured results under two conditions that we arbitrarily set up. We ask such questions as to whether the eye is more sensitive during stimulation of other sense organs or in the absence of such stimulation; whether boys or girls are more capable in a test of perceptual speed; whether one method of teaching subtraction is superior to another in terms of resulting efficiency. This calls for one set of measurements under the one condition and another set under the other condition and a comparison of means. The statistical question is, "How reliable is the difference between means?"

The Standard Error of a Difference between Uncorrelated Means.— Again reliability is indicated by a standard error. The amount of fluctuation in a difference between sample means is naturally related to the amount of fluctuation in the means themselves. The simplest relationship is given by the formula

$$\sigma_{d_M} = \sqrt{\sigma^2_{M_1} + \sigma^2_{M_2}} \quad \text{(Standard error of a difference between uncorrelated means)} \qquad (9.29)$$

where σ_{M_1} = SE of the mean of the first distribution.
σ_{M_2} = SE of the mean of the second distribution.

This relationship holds only when the two sets of measurements are independent, i.e., uncorrelated. When we are dealing with matched groups, for example, particularly when individuals are matched pair by pair, the formula will have to be enlarged. But more of that later.

Let us apply formula (9.29) to a typical problem. A group of 114 men and a group of 175 women were given the same word-building test in which the score is the number of words built out of six letters in 5 min. The results are given in summarized form in Table 9.5. The women's mean of 21.0 is 1.3 points higher than that for the men. This mean difference is very small numerically, but in view of the relatively large number of cases in the two samples, we should expect the obtained means to be very close to the true means, and perhaps therefore it indicates a

TABLE 9.5.—MEANS AND OTHER STATISTICS IN THE COMPARISON OF MEN AND WOMEN IN A WORD-BUILDING TEST

Statistic	Men	Women
N	114	175
M	19.7	21.0
σ	6.08	4.89
σ_M	.572	.371
σ_{d_M}	.682	
D_M	1.3	
t	1.91	

real sex difference. The stability of each mean is indicated by its *SE*, which is .572 in the case of the men and .371 in the case of the women.

Just as sample means are distributed normally about the true mean when N is large, the sample differences between means are also distributed normally. The central tendency about which the differences between means fluctuate is a population value. We do not know what that population value is. We are most concerned, first, in determining whether there is any difference at all, and second, in determining its approximate size. The statistical tests connected with differences, in principle, are very much like those we encountered in connection with correlation coefficients. Since most differences are small (if they were not, we should hardly need to make statistical tests) we first make a test to see whether we are justified in rejecting the null hypothesis. The null hypothesis in this case is the supposition that in the population there is no real difference. Stated in another, and more acceptable, way, the null hypothesis is that the two sample means arose by random sampling from the *same* population. Same, that is, with respect to the variable measured; the two groups from which the two samples were drawn are obviously different in other respects, otherwise we would not have raised any question of a difference at all.

In accordance with the null hypothesis, then, we assume a sampling distribution of differences, with the mean at zero, or at $\tilde{M}_1 - \tilde{M}_2 = 0.0$. The deviation of each sample difference, $M_1 - M_2$, from this central reference point is equal to $(M_1 - M_2) - (\tilde{M}_1 - \tilde{M}_2)$, or $M_1 - M_2 - 0$. The deviation of each difference given in terms of standard measure would be the deviation divided by the standard error, which gives us a t value. In terms of a formula,

$$t = \frac{M_1 - M_2}{\sigma_{d_M}} \qquad \text{(A } t \text{ ratio for a difference between means)} \qquad (9.30)$$

The numerator, to be quite complete, should read $M_1 - M_2 - 0$, as was stated above, but since the zero has no contribution to make to the computation, it is dropped in ordinary practice. It will help the investigator using this formula to think more clearly if he remembers that logically the zero belongs there.

Figure 9.5 shows graphically a sampling distribution of t ratios. This distribution is real, though rarely derived by using actual data, because every difference we obtain by random sampling, with N's constant, provides its own t value. We could actually take a series of 100 paired samples, compute $M_1 - M_2$ for each pair, σ_{d_M} for each pair, and conse-

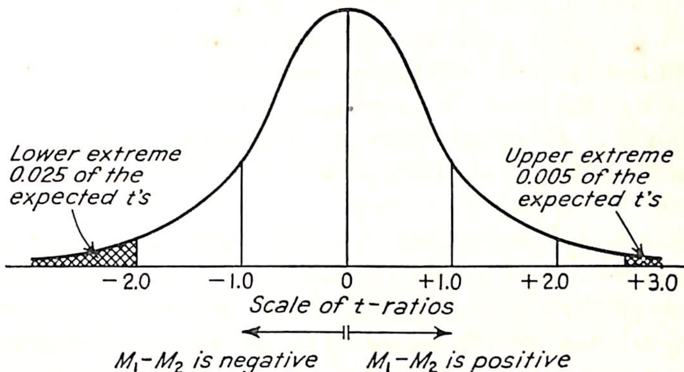


FIG. 9.5.—A sampling distribution of t with a mean of 0, which corresponds to a hypothetical difference between means equal to zero. Shaded areas show the regions of extreme t's; at the left those significant at the 5 per cent level and at the right those significant at the 1 per cent level. Obtained t's (*either positive or negative*) in those extreme regions are interpreted accordingly.

quently a t. The frequency distribution of the 100 t's we could set up from those data would look like Fig. 9.5.

Testing the Null Hypothesis.—For the word-building test we have the information (see Table 9.5) that the difference $M_1 - M_2$ is -1.3. The algebraic sign of the difference does not concern us at this time; we are interested only in its amount. The standard error $\sigma_{d_M} = .682$. From this,

$$t = \frac{1.3}{.682}$$
$$= 1.91$$

The value 1.91 tells us how many σ_{d_M}'s the obtained difference extends from the mean of the distribution. The mean, under the null hypothesis which is being tested, is a difference of zero. Since the sample is large, we may assume a normal distribution of the t's and interpret the obtained t accordingly. It fails by just a little to meet the 5 per cent level of confi-

dence (which for large samples is 1.96); consequently we would not reject the null hypothesis and we would say that the obtained difference is not significant. There may actually be some difference, but we have not enough assurance of it. There are more than 5 chances in 100 that a difference as large as this one, or larger, could have happened by random sampling from the same population—same with respect to word-building ability. A more practical conclusion would be that we have insufficient evidence of any sex difference in word-building ability, at least in the kind of population sampled. Note that the conclusion was *not* stated to the effect that we have demonstrated that there is *no* sex difference in word-building ability. *We cannot prove the truth of the null hypothesis; we can only demonstrate its improbability.*

Had the t test turned out very significant, *i.e.*, with less than 1 chance in 100 that by chance a t could be so large, we would then have been interested in the *size* of the difference. Our interest would then have reverted to the standard error of the difference and the probable limits it suggested for the size of the difference. This procedure is so similar to that for determining the probable size of any population parameter that we need not go through the steps here.

The Standard Error of a Difference in Correlated Data.—When the data are so sampled that there is a correlation between the means in the two variables measured, *i.e.*, so that the means in pairs of samples tend to rise or fall together (positive correlation) or tend to be contrasting so that when one rises the other falls (a negative correlation), the SE of a difference is estimated by the formula

$$\sigma_{d_M} = \sqrt{\sigma^2_{M_1} + \sigma^2_{M_2} - 2r_{12}\sigma_{M_1}\sigma_{M_2}} \quad \text{(Standard error of a difference between correlated means)} \quad (9.31)$$

which is like formula (9.29) except for the last term, in which r_{12} is the correlation *between the two sets of means.*

Fortunately, under the usual circumstances of random sampling, the correlation between the two sets of means is approximately equal to the correlation between two sets of single measurements in two samples. Since we ordinarily have only two samples with two means from which we could not compute r_{12}, this fact is a great convenience. But in order to compute the correlation between single measurements, we must have the individual measurements in the two samples paired off two by two in some manner. For example, if the same group of students takes the same word-building test twice instead of two different groups taking it, we have the same individual's score in the first trial to pair off with his score in the second trial. Or if, in comparing males and females in the

test, we want to standardize our two groups better by taking a brother and a sister from each family or if we pair boy with girl with respect to age, *IQ*, or social status, or all such factors, then if these factors of common family, common age, *IQ*, or social status have any relation to word-building score, they automatically introduce correlation into the two samples. We compute a coefficient of correlation in the manner described in Ch. 8 and introduce it into formula (9.31).

In Table 9.6, we find two sets of knee-jerk measurements, both from the same 26 men but under two conditions. In the first case (*T*), the subjects were squeezing a hand dynamometer just before the stimulus struck the knee, and in the second case (*R*) the "relaxed" knee jerk was obtained under a relaxed, sitting posture. Will the average man show a real difference in height of knee jerk under the tensed condition, as theory would lead us to expect? The two means, with a difference of 3.39 deg., suggest that the theory is vindicated. But we want to be sure that this large a difference could not have happened by random sampling from a population of measurements in which the actual difference is zero.

If we were to assume no correlation between the tensed and normal measurements of knee jerk, we should apply formula (9.29), or we should apply formula (9.31) with an r_{12} equal to zero, which is actually the same thing. Such a σ_{d_M} turns out to be 2.37 deg. of arc. The t ratio is 3.39/2.37, or 1.43. This t falls decidedly short of the 5 per cent level of significance. We should conclude, erroneously, that although there is some difference in the expected direction, it is not a significant one. So far as these indications go, we would not be called upon to reject the null hypothesis; the difference of 3.39 could represent merely a result of random sampling.

When we compute a coefficient of correlation between the two sets of measurements, we find it to be +.82. This means that the men came rather closely in the same rank order in both the tensed and the relaxed conditions. If a man has a high kick under normal conditions, he will be likely to have a correspondingly high kick during the tensed conditions. If a man is low in the one case, he is likely to be low in the other. If the sampling is random, there would be a similar correlation between *means* under the two conditions. If another group of 26 men had a higher normal average response than this one, it would be likely also to have a higher average tensed response. When means rise and fall together, they tend to maintain the same difference between them. In the case of a perfect positive correlation ($r = +1.0$), the difference between means would remain exactly constant. If all the sample differences between means were identical, their dispersion would be zero, and, σ_{d_M} would equal

Table 9.6.—Strength of the Patellar Reflex under Two Conditions, Tensed and Relaxed, for 26 Men, and Differences between Them (Measurements Are in Terms of Degrees of Arc)

	T Tensed	R Relaxed	T − R Difference
	31	35	− 4
	19	14	+ 5
	22	19	+ 3
	26	29	− 3
	36	34	+ 2
	30	26	+ 4
	29	19	+10
	36	37	− 1
	33	27	+ 6
	34	24	+10
	19	14	+ 5
	19	19	0
	26	30	− 4
	15	7	+ 8
	18	13	+ 5
	30	20	+10
	18	1	+17
	30	29	+ 1
	26	18	+ 8
	28	21	+ 7
	22	29	− 7
	8	4	+ 4
	16	11	+ 5
	21	23	− 2
	35	31	+ 4
	26	31	− 5
Σ	653	565	+88
M	25.12	21.73	3.39
σ	7.17	9.45	5.50
σ_M	1.43	1.89	1.10

zero. We would then be almost certain of a true difference in the obtained direction. A correlation of +.82 is less than 1.00, however; so there is still some room for variability among the differences. But from the line of reasoning just completed, we can see that the σ_{d_M} is going to be smaller than it turned out to be when we assumed an r equal to zero.

By the use of the complete formula (9.31), we find the σ_{d_M} to be 1.10, which is less than half the previous estimate of 2.37. The t ratio is now $3.39/1.10 = 3.06$. A t above 3 is obviously in the "very significant" category.[1]

We therefore feel very confident that there is a real difference in favor of the tensed conditions. This is not saying that we feel sure that the true difference is exactly 3.39; it might be more or less than that. At any rate, the hypothesis with which the experiment started receives substantial support from the result.

Observations Should Often Be Paired.—In setting up an experiment with two groups of subjects or two groups of measurements for statistical comparison, it is well to pair off cases two by two if possible, so that a correlation can be computed. Often when such pairing is not actually carried out, there would still be correlation between means of samples anyway; the full formula for the *SE* of a difference cannot then be applied, and the σ_{d_M} by formula (9.29) is overestimated. It is true that under these circumstances, if the correlation is positive, as is usually the case when there is correlation, we can say that the correct σ_{d_M} is smaller and that the correct t ratio is larger than the one we estimated. When we have a significant or very significant t ratio under these circumstances, we can be sure that the t we would obtain by taking into account the positive correlation would be even larger. But one difficulty is that when the t ratio obtained under these circumstances is too small to be significant, we cannot conclude anything in particular. Least of all can we conclude that the true difference is probably zero, for had we considered the correlation, we might have found a significantly large t ratio. The process of matching and the inclusion of the correlation factor in the σ_{d_M} formula are said to increase the *precision of the t test.* By this is meant that the test is more sensitive to a difference when it is real. As a result, we are more likely to avoid the error of accepting the null hypothesis when it is incorrect.

In pairing off individuals or observations, it is important that the pairing be done on some significant basis. It will not pay to do any pairing except on the basis of some trait that correlates with the measurements on which the two groups are going to be compared. For example, if we

[1] A sample of 26 pairs of observations would be regarded as a "small sample" by many investigators. In this case, a larger t would be required for meeting the confidence levels of significance (see Table 9.4). In this problem, with 25 degrees of freedom, a t of 2.79 is significant at the 1 per cent level (see Table D). The obtained t (3.06) also exceeds this limit. We have $N - 1$ degrees of freedom in matched data, where N is the number of *pairs* of observations.

were to compare two groups of boys as to ability to do a high jump, one group after training of a certain kind and the control group without such training, it would be important that the two groups be equated as to age, among other things. Ability in the high jump, regardless of training, would be dependent upon age, hence correlated with it. But the ability is probably not correlated significantly with grade earned in arithmetic; so there would be no point in matching the groups on this variable.

The basis upon which to match groups having been decided, there are two common ways of carrying out the matching. One is by pairing cases directly. In the problem just mentioned, for every boy of ten years six months in the one group, one would seek a boy of like age in the other. Small discrepancies may well be permitted at times between pairs. If there are about twice as many cases in the one sample as in the other, matching two boys to one would be the solution. The other common way of matching groups is to ignore individuals as such and simply to try to make sure that the two samples have approximately equal means, standard deviations, and skewness. When this is done and the two variables are correlated, the formula for the standard error is[1]

$$\sigma_{d_M} = \sqrt{(\sigma^2_{M_1} + \sigma^2_{M_2})(1 - r^2_{mx})} \quad \text{(Standard error of a difference for matched samples)} \qquad (9.32)$$

in which r_{mx} is the correlation between X_m (the variable on which the groups were matched) and X (the variable on which we are testing the difference). If the groups are matched on the basis of two or more variables, a multiple correlation coefficient is involved (see Ch. 16).

Comparison of formula (9.11) with formula (9.32) will show that they are alike. The former corrects σ^2_M for the effects of matching while the latter corrects simultaneously the two variances of means that enter into the formula for σ_{d_M}. This should be sufficient warning that if the correction has previously been made in each σ^2_M, the correction given in formula (9.32) should not be used; that would be applying the same correction twice.

A Standard Error of a Difference Obtained Directly from Differences.—When individuals have been paired off, we can find the desired statistics directly from differences between pairs. In Table 9.6, we find the difference in knee-jerk measurements ($T - R$), given with algebraic signs, for every individual. If we sum them and divide by N, we obtain the mean of the differences, which is equal to the difference between the means. If we calculate the *SE* of the mean of these differences, we have σ_{d_M}. The σ_{d_M} is thus obtained in the most direct manner. We do not even need to

[1] McNemar, Q, *Psychol. Bull.* 1940, **37**, 331–365.

know the *SE*'s of the two means or the amount of correlation present, yet our direct procedure has taken these things into account. The σ_{d_M} for the knee-jerk data obtained in this manner is identical with that which we found previously, as it should be. The interpretations and conclusions concerning the mean difference are the same as usual. This more direct method is very strongly recommended whenever it can conveniently be applied.

The Reliability of Differences between Proportions, Frequencies, and Percentages.—Consider the data in Table 9.7. Here we have the proportions of 400 men and of 400 women students who judged two words as "pleasant" or "very pleasant." The two words were "to explore" and "symphony." Here we can raise two questions concerning each word. Is there any sex difference in the proportion judging the word "pleasant"? And within each sex, is there a significantly greater proportion of "pleasant" judgments for one word than for the other? The differences themselves show that the men favor the word "to explore" slightly more than

TABLE 9.7.—PROPORTIONS OF 400 MEN AND 400 WOMEN WHO JUDGED THE WORDS "TO EXPLORE" AND "SYMPHONY" PLEASANT; DIFFERENCES AND STANDARD ERRORS OF DIFFERENCES; AND t RATIOS

	"to explore"	"symphony"	r	Difference	σ_{d_p}	t
Men	.8775	.6850	.342	.1925	.0234	8.23
Women	.8700	.8875	.395	.0175	.0180	0.97
Difference	.0075	.2025				
σ_{d_p}	.0235	.0281				
t	0.32	7.21				

do the women, the difference in proportion being .0075. The women decidedly more often favor the word "symphony," with an excess of .2025 over the proportion of the men who judge it pleasant. The men find the word "to explore" more pleasing than they do the word "symphony" by a margin of .1925, and the women, on the other hand, find the word "symphony" more to their liking than "to explore" by a small margin of .0175. Which of these differences, if any, are significant or very significant according to the rules we have been following? We can test any or all of them for statistical significance.

The Standard Error of a Difference between Proportions.—The standard error of a difference between two proportions is given by the formula

$$\sigma_{d_p} = \sqrt{\sigma^2_{p_1} + \sigma^2_{p_2} - 2r_{12}\sigma_{p_1}\sigma_{p_2}} \quad \text{(Standard error of difference between proportions)} \quad (9.33)$$

where σ_{p_1} = SE of the first proportion.

σ_{p_2} = SE of the second proportion.

r_{12} = correlation of proportions in pairs of samples.[1]

Again, it is fortunate for us that, when sampling is random, the correlation between proportions is equal to the correlation between single cases. The latter we can estimate from the data. In Table 9.7, we find that the correlation between men's judgments of the two words is given as +.342 and the correlation for the women is +.395, since both words were judged by the same individuals. But in the comparison between sexes, there was no pairing of individual judgments in any known way; so we may assume that the correlations are zero. On this basis we find the σ_{d_p} between men and women for the word "to explore" to be .0235. The obtained difference of .0075 here yields a t ratio of 0.32, which is decidedly not significant. The sex difference on the word "symphony" gives a σ_{d_p} of .0281, which yields a t ratio of 7.21. This is so far above the limit for "very significant" deviations that we are very confident about its being true that college women (like those in the sample) find "symphony" more pleasant than do college men (like those in the sample). Men also decidedly prefer "to explore" to "symphony," with the highly significant t value of 8.23. Women, however, who find "symphony" more pleasing than "to explore" by an excess of .0175, do not give any sure indication that the true difference is in this direction, for the t ratio is only 0.97. The results are somewhat in line with what we should expect, but it can be ventured that some differences that we expected to be true did not prove to be significant and perhaps do not exist at all; for example, where we might have expected a difference between sexes on "to explore," a significant one failed rather decisively to appear.

Differences between Percentages and Frequencies.—Similar tests of significance can be made for differences between percentages and frequencies. The uses of percentages and frequencies are here completely analogous to the use of proportions as they have been in other connections. An illustration of how to test either of these differences will therefore not be given.

The Reliability of Differences between Standard Deviations.—If we are concerned about differences in variability in two distributions as measured by σ, we can also make statistical tests of significance somewhat like the ones already illustrated. The formula for the standard error of a difference between σ's is

$$\sigma_{\sigma_d} = \sqrt{\sigma^2_{\sigma_1} + \sigma^2_{\sigma_2} - 2r^2_{12}\sigma_{\sigma_1}\sigma_{\sigma_2}} \quad \text{(Standard error of a difference between standard deviations)} \qquad (9.34)$$

[1] This correlation should be derived from samples as a ϕ coefficient, or the correlation of two genuinely dichotomous variables (see Ch. 13).

It is especially to be noted that the r_{12} in this equation, unlike its appearance in others, is squared, for it has been proved that the correlation between standard deviations in pairs of samples is equal to the square of the correlation coefficient between individual pairs of measurements; hence the squaring in formula (9.34).

We may apply this formula to the data in Table 9.5 for the word-building test. Here we find the men more variable than the women by a difference of 6.08 − 4.89, or 1.19 points. Is this difference significant, or could it have arisen as a natural deviation from an actual difference of zero, *i.e.*, equality of the sexes in variability? The σ_{d_σ} proves to be .476 (the correlation being zero) and the t ratio is 1.19/.476, or 2.50. The difference of 1.19 points therefore just fails to pass the hurdle of significance at the 1 per cent level. There is just more than one chance in a hundred that if the two sexes are equally variable in this test, such a large discrepancy between their standard deviations could have occurred by sampling. Just failing to "pass the hurdle," however, should not be stressed too much. The amount of difference obtained is a very rare occurrence and strongly suggests the inference that there is a real sex difference in variability in the word-building test.

Under the heading of small-sample statistics will be found a radically different method for testing a difference between two standard deviations. With small samples the test given above breaks down completely for lack of normal sampling distributions.

Reliability of Differences between Coefficients of Correlation.—If we have two coefficients of correlation, r_{12} and r_{34}, that have been obtained from intercorrelating two pairs of variables and we want to test whether they could have arisen from the "same population" by random sampling, by analogy to other formulas, the standard error of a difference between r's is estimated by

$$\sigma_{d_r} = \sqrt{\sigma^2_{r_{12}} + \sigma^2_{r_{34}} - 2r_{r_{12}r_{34}}\sigma_{r_{12}}\sigma_{r_{34}}} \qquad (9.35)$$

(Standard error of the difference between two coefficients of correlation with no common variable)

where $\sigma_{r_{12}}$ = the standard error of r_{12}.
$\sigma_{r_{34}}$ = the standard error of r_{34}.
$r_{r_{12}r_{34}}$ = the correlation between samples or r_{12} and r_{34}.

The estimation of the correlation of r's can be made by means of a very long formula involving r_{13}, r_{14}, r_{23}, and r_{24}, as well as r_{12} and r_{34}, which makes this procedure forbidding. With no variable in common to the two r's being compared, it is likely that the r between r's will be rather small. When one of the variables in the r_{12} correlation is very highly

correlated with one in the r_{34} correlation, however, the r_{rr} correlation would probably be of sufficient size to call for its use.

The type of problem in which the average reader will be likely to test differences between r's is one in which one of the variables is common to the two correlations. This calls for a different correlation of correlations (see formula 9.36). For this reason the reader is referred elsewhere for the method of estimating $r_{r_{12}r_{34}}$.* Without using the correlation term r_{rr}, one can sometimes reject the null hypothesis with confidence, because t is underestimated, but sometimes one could not feel very sure that he should *not* reject it if r_{rr} is of substantial size and is not used.

In experimental investigations in which we study the change in correlation (perhaps reliability or validity) of a measuring instrument under different conditions, one or both of the correlated variables is likely to enter into both correlations. We determine the validity correlation for a test with and without scoring weights using the same outside criterion. We compare the validity coefficients of two similar verbal tests, also against the same criterion. For such a situation we would be testing the difference between two correlations r_{12} and r_{13}, where variable X_1 is common to both. If we substitute r_{13} for the correlation r_{34} in formula (9.35), we can estimate the standard error σ_{d_r} for these two correlations. The correlation of the r's would be $r_{r_{12}r_{13}}$. This correlation can be estimated by the formula

$$r_{r_{12}r_{13}} = r_{23} - \frac{r_{12}r_{13}(1 - r^2_{23} - r^2_{12} - r^2_{13} + 2r_{12}r_{13}r_{23})}{2(1 - r^2_{12})(1 - r^2_{13})} \tag{9.36}$$

(Correlation between two r's having one variable in common) (9.36)

The **z** *Test of Differences between r's.*—Remembering that there are doubts about the use of standard errors of r's when correlations are large and when samples are not large, it would be well to consider testing differences between **z** coefficients instead. Unfortunately, no one appears to have found a way of estimating correlations between paired samples of **z**'s. We must therefore be limited to problems in which r_{zz} is very small or zero: as when the two correlations being compared arose from rather independent variables.

With this limitation, the standard error of a **z** difference is

$$\sigma_{d_z} = \sqrt{\frac{1}{N_1 - 3} + \frac{1}{N_2 - 3}} \tag{9.37}$$

(Standard error of a difference between two **z** coefficients) (9.37)

* Peters, C. C., and Van Voorhis, W. R. Statistical procedures and their mathematical bases. New York: McGraw-Hill, 1940. P. 185.

Consider two r's, $r_{12} = .82$ and $r_{13} = .92$. The corresponding **z** coefficients (from Table H) are 1.16 and 1.59. $N_1 = 50$ and $N_2 = 60$, from which

$$\sigma_{d_z} = \sqrt{1/47 + 1/57} = \sqrt{.03882} = .197$$

The t ratio is equal to

$$\frac{1.59 - 1.16}{.197} = \frac{.43}{.197} = 2.18$$

From this we would feel more confident than usual that the difference is significant at the .05 level or better, for had we taken into account a possible positive correlation between **z**'s, the t would have been larger.

Small-sample Statistics

The distinction between large-sample and small-sample statistics is not an absolute one, by any means, the one realm merging into and overlapping so extensively the other. If one asks, "How small is N before we have a small sample?" the answers from different sources will vary. There is general agreement that the division, if there must be one, is in the range of 25 to 30. Some place it as low as 20 and others say that anything under 100 is a small sample. The truth of the matter is that the needs for small-sample considerations increase as N decreases and they may become critical somewhere below an N of 30. Sampling distributions depart from the normal form more and more as N decreases. This was first realized by W. S. Gosset, who published for many years under the mysterious name of "Student," and it was later emphasized by R. A. Fisher, who has worked out many of the procedures.

The Sampling Distribution of t.—For small samples, many statistics exhibit sampling distributions that depart from normality in various ways, as was indicated in connection with discussions of standard errors in earlier sections of this chapter. Distributions of correlation coefficients, proportions, and of standard deviations are often skewed. Another important change that affects distributions of differences particularly is a change in *kurtosis*. Kurtosis is apparent in the degree of "peakedness" of the center of the distribution. A normal distribution is called *mesokurtic*, which means neither very peaked nor very flat across the top. Curves tending toward rectangular form, more or less, are called *platykurtic*. Those more

peaked than normal are called *leptokurtic.* The distribution of t tends to be leptokurtic. Figure 9.6 shows a leptokurtic distribution compared with a normal one. The most important thing to notice is not the sharpness of the center but the fact that the tails of the leptokurtic curve are higher than for the normal curve. The greater areas under the two tails mean

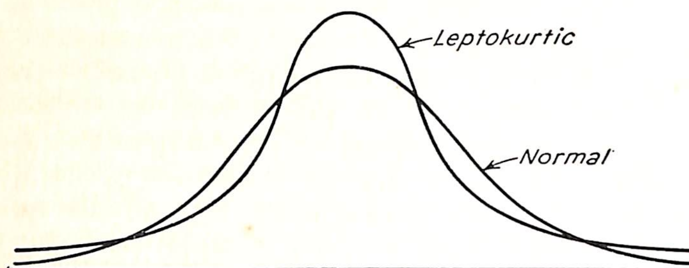


FIG. 9.6.—Comparison of a normal distribution with a leptokurtic distribution when their means and standard deviations are approximately equal.

that we would have to go out to greater deviations in terms of σ units in order to include the same proportion of area inside the limits of those deviations. If we ask how many units one must go from the mean in both directions to include all except .05 of the area, the answer would be larger for the leptokurtic than for the mesokurtic distribution.

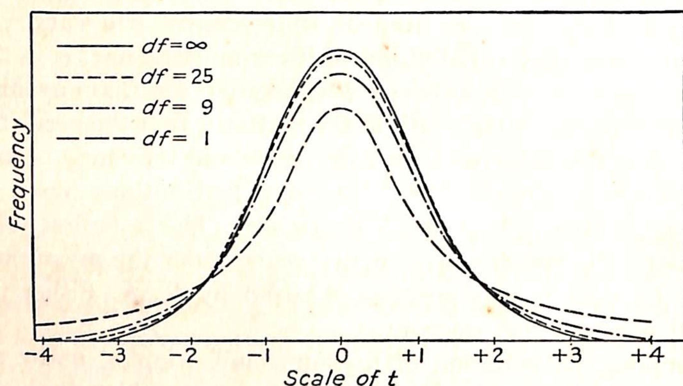


FIG. 9.7.—Student's sampling distribution of t for various degrees of freedom. As the *df* become infinite, the distribution of t becomes normal. (*After Lewis, D. Quantitative methods in psychology. Iowa City. The author,* 1948.)

The smaller the number of degrees of freedom, the farther the kurtosis shifts from the normal form. This is shown by Fig. 9.7. With a very large number of degrees of freedom we have the normal curve. With 25 degrees of freedom the departure from normality is so slight, except very near the mean (which does not matter for the t test) and at the extreme tails, that we usually would not go far wrong in assuming normality.

With 9 and 1 degrees of freedom, however, the t distributions depart drastically from mesokurtosis. The figure lends support to the choice of 25 as a lower limit to large-sample logic and practice.

Confidence Limits in the t Distribution.—Refer again to the high tails of the t distribution with small samples, in Fig. 9.7. The t values required for significance at the .05 and .01, and other levels of confidence have been calculated. For the .05 and .01 levels the required t's are given in Table D, last column. For very large samples, the two t's are 1.960 and 2.576, respectively. For a sample of 1,000 *df* the limits change in the third decimal place only. For 100 *df* there is a little change in the second decimal place. The limits with 100 *df* are 1.984 and 2.626. Rough limits, by rounding, of 2.0 and 2.7 would do very well even down to about 30 degrees of freedom. With only 10 *df*, however, t's of 2.23 and 3.17 would be required for the respective confidence levels. One could, of course, look up the t's required for significance to suit the number of *df* in his particular investigation. With small samples this becomes imperative, if one is to make the proper inferences from t tests.

Fisher's *t* Formulas.—Fisher has provided several formulas designed for the computation of t when samples are small. We will first note his t formula in connection with a coefficient of correlation.

For the Test of a Coefficient of Correlation.—If the population r is zero, or may be assumed to be zero as when we assume the null hypothesis for the purpose of testing it, the t desired for this test is estimated by the formula

$$t = r\sqrt{\frac{N-2}{1-r^2}} \quad \text{(Fisher's formula for } t \text{ in testing a coefficient of correlation)} \qquad (9.38)$$

where r = obtained coefficient of correlation

N = number of pairs of observations from which r was computed.

Applying this to an illustrative problem we considered earlier, where $r = .30$ and $N = 50$,

$$\begin{aligned} t &= .30\sqrt{\frac{48}{.91}} \\ &= .30\sqrt{52.75} \\ &= (.30)(7.26) \\ &= 2.18 \end{aligned}$$

We may therefore regard the obtained correlation as probably not representing a population correlation of zero, though we can reject the null hypothesis just beyond the 5 per cent level of confidence. According to Table D, with the 48 *df* we have here, the two required t's are 2.01 and 2.68.

For the Test of a Difference between Means.—When means are uncorrelated, the t formula for testing their difference is

$$t = \frac{M_1 - M_2}{\sqrt{\left[\frac{\Sigma x^2_1 + \Sigma x^2_2}{N_1 + N_2 - 2}\right]\left[\frac{N_1 + N_2}{N_1 N_2}\right]}} \quad \text{(Fisher's } t \text{ formula for testing the difference between means)} \quad (9.39)$$

where M_1 and M_2 are the means in the two samples.
Σx^2_1 and Σx^2_2 are the sums of squares in the two samples.
N_1 and N_2 are the numbers of observations, respectively.

The numerator is identical with that in formula (9.30), and as in that place, it should read $M_1 - M_2 - 0$, if it were written in full to represent the deviation that it is. The denominator as a whole is the standard error of the difference between means, as any t ratio requires, but no doubt it appears very unfamiliar. In writing the σ_{d_M} in this form, Fisher has taken the null hypothesis quite seriously, as we *should* do if we are completely consistent. That is, if there is but *one* population there should be but *one* estimate of its variance. The variances in formula (9.29) are allowed to differ, as they come from two different samples. But if they came from the *same* population, any difference between them should be due merely to sampling errors. The first term under the radical in formula (9.39) is a single estimate of the population variance. The numerator of the term sums the sums of squares coming from the two samples. This is divided by the total number of degrees of freedom, which for this problem is $N_1 + N_2 - 2$. The same *df* could be found by summing $(N_1 - 1) + (N_2 - 1)$. The use of the second term under the radical has the effect of computing the variance of the mean of differences from the estimated population variance.

When the two samples are of equal size, *i.e.*, $N_1 = N_2$, formula (9.39) simplifies to

$$t = \frac{M_1 - M_2}{\sqrt{\frac{\Sigma x^2_1 + \Sigma x^2_2}{N_i(N_i - 1)}}} \quad \text{(} t \text{ ratio for difference between means in two samples of equal size)} \quad (9.40)$$

where N_i = size of either sample.

When means of paired samples are not independent but correlated, the best formula to use for deriving t directly from sums of squares is

$$t = \frac{M_d}{\sqrt{\frac{\Sigma x^2_d}{N(N - 1)}}} \quad \text{(The } t \text{ for differences between correlated pairs of means)} \quad (9.41)$$

where M_d = mean of the N differences of paired observations.
x_d = deviation of a difference from the mean of the differences.
The procedure implied by this formula was actually applied earlier in connection with the knee-jerk data under two experimental conditions (see Table 9.6). The number of degrees of freedom to use in this case is $N - 1$, where N is the number of *pairs* of observations. For the knee-jerk problem there are 25 degrees of freedom, which indicate t's of 2.06 and 2.79 for the .05 and .01 levels, respectively.

Differences between Means in Nonnormal Distributions.—If there is good reason to believe that the population distribution is not normal but seriously skewed or bimodal, and especially if the samples are small, the usual t test does not apply. For such a situation, methods developed by Festinger, and others, are probably the most suitable substitutes.[1] Such problems are not sufficiently common to justify explaining those methods here.

For the Test of a Difference between Uncorrelated Proportions.—When the null hypothesis is assumed with regard to two observed proportions, Fisher recommends, again, that we use just one estimate of the population variance. This requires the use of a weighted mean of the two sample proportions. Formula (4.8), previously given, can be applied here. It is repeated here to apply to the averaging of two proportions.

$$\tilde{p}_e = \frac{N_1 p_1 + N_2 p_2}{N_1 + N_2} \quad \text{(A weighted mean of two sample proportions used to estimate a population proportion)} \tag{9.42}$$

The formula for t is

$$t = \frac{p_1 - p_2}{\sqrt{\tilde{p}_e \tilde{q}_e \left(\frac{N_1 + N_2}{N_1 N_2}\right)}} \quad \text{(A } t \text{ ratio for a difference between uncorrelated proportions)} \tag{9.43}$$

where $\tilde{q}_e = 1 - \tilde{p}_e$.

When the two samples are of equal size, *i.e.*, $N_1 = N_2$, if we let both equal N_i, formula (9.43) simplifies to

$$t = \frac{(p_1 - p_2)}{\sqrt{\frac{2\tilde{p}_e \tilde{q}_e}{N_i}}} \tag{9.44}$$

Since this formula is proposed as a "small-sample" device, the question may arise as to just how small a sample will be suitable for the application

[1] Festinger, L. The significance of difference between means without reference to the frequency distribution function. *Psychom.*, 1946, **11,** 97–105.

of the formula. Statisticians seem to have very little to say on this point or on the question of degrees of freedom in connection with this particular t test. It would seem to the author that we should be cautious about applying formula (9.43) to very small samples for the same reason, as was earlier stated, that the use of σ_p is of dubious validity when a small sample is combined with an extreme proportion. Particularly to be avoided is the application of this formula when p_1 or p_2 is 0.0 or 1.0.

Differences between Correlated Proportions.—While formula (9.33) is general enough to take care of testing the significance of differences between proportions (also percentages and frequencies), in most instances when data are correlated there is a more economical procedure recently introduced by McNemar.[1] As stated in an earlier footnote, the correlation required by formula (9.33) is the ϕ coefficient or the correlation coefficient with two categories in both X and Y. McNemar's formula avoids the necessity for computing the standard errors of the proportions as well as the phi coefficient of correlation, but does require having the data in a form that *can* be used for computing ϕ.

For a genuine nonzero correlation to exist between the two samples, as usual, either the same individuals or objects must appear in both or there must be a pairing in some significant manner, as of twins, siblings, or experimental-control pairs. Suppose we have administered two test items to a sample of 100 students. Item I is answered correctly by 60 of the group and item II by 70. Is item II actually easier than item I? In making the t test to answer this question, we must definitely face the possibility of correlation between the two items and consequently between the two proportions. To handle this problem properly, we need to set up the data in the form of a four-cell contingency table, as in Table 9.8. At the left are the four frequencies of those who were correct on item I and either correct or incorrect on item II, and the frequencies of those who were incorrect on item I and either correct or incorrect on item II. At the right in Table 9.8 are given letter symbols to stand for the four categories. Using these symbols, McNemar's formula, in modified form, reads

$$t = \frac{b - c}{\sqrt{b + c}} \quad \text{(t ratio for difference between correlated proportions)} \qquad (9.45)$$

It will help to assure the proper application of this formula to note that the symbols b and c stand for the discordant cases in the four-cell table; in this problem b and c stand for individuals who succeed in one item and

[1] McNemar, Q. Note on the sampling error of the difference between correlated proportions or percentages. *Psychom.* 1947, **12,** 153–157.

TABLE 9.8.—A FOUR-CELL CONTINGENCY TABLE OF FREQUENCIES OF STUDENTS WHO PASSED OR FAILED EACH OF TWO TEST ITEMS

Frequency Table

Item I	Item II: Fail	Pass	Both
Pass	5	55	60
Fail	25	15	40
Both	30	70	100

Symbolic Table

Item I	Item II: Fail	Pass	Both
Pass	b	a	$a + b$
Fail	d	c	$c + d$
Both	$b + d$	$a + c$	N

fail in the other. It will also help to know that the difference $b - c$ divided by N equals the difference between p_1 and p_2. It is therefore the difference between two obtained *frequencies*, *i.e.*, $b - c = Np_1 - Np_2$. To find the difference that is being tested in the numerator of the t ratio is not a new experience. The denominator, therefore, must somehow represent the standard error of a difference between frequencies (it would be N times the standard error of a difference between proportions) *with the correlation taken into account*. In this formula, too, there is implied but one estimate of the population variance and it is derived from an average of the sample proportions.

Solving formula (9.45) as applied to the test-item data, we have

$$t = \frac{5 - 15}{\sqrt{5 + 15}} = \frac{-10}{\sqrt{20}} = \frac{-10}{4.47} = -2.24$$

The difference we would infer to be significant between the .05 and .01 levels. Item II is probably easier than item I.

It is informing to see what the outcome would have been if we had applied formula (9.44), without taking into account the amount of intercorrelation. With $\bar{p}$ estimated to be .65,

$$t = \frac{.10}{\sqrt{\frac{2(.65)(.35)}{100}}} = \frac{.10}{\sqrt{\frac{.4550}{100}}} = \frac{.10}{.06745} = 1.48$$

From this result we would have concluded that the difference was insignificant. This demonstrates how a decision may be altered drastically when the correlation term in the standard-error formula is taken into account. Without it, we run the risk of making an error of the second kind; of not rejecting the null hypothesis when it is false. The correlation (ϕ coefficient) between the two items amounts to +.58. The reader will find that if he lets $\sigma^2_{p_1} = \sigma^2_{p_2} = \sigma_{p_1}\sigma_{p_2} = .002275$, and substitutes these with the correlation of +0.58 in formula (9.33) he will come out with a σ_{d_p} equal to .0439, which gives a t of 2.29, which is near that obtained with McNemar's formula (2.24).

One restriction in the application of formula (9.45) is that $b + c$ should be 10 or greater.

The F Test of Differences between Standard Deviations.—For small samples, the t test of differences between standard deviations is not satisfactory, even with the availability of Student's distribution for t. Instead of testing the significance of a difference between two σ's, we can test the significance of the *ratio of the two variances* that correspond to them. If we compute the ratio of the larger of two variances to the smaller of the two, the larger the difference, the further the ratio exceeds 1.00. The ratio is 1.00 when the two variances are equal. If the ratio of the variances is significant, the difference between the standard deviations is significant.

More accurately stated, we do not find the ratio of the variances in the two samples. Instead, we find an estimate of the population variance from each of the two random samples and from these values compute the ratio. We assume the null hypothesis, that the two samples came from the same population, and we ask whether two estimates of that population variance could differ as much as the obtained ratio indicates. The ratio has been given the symbol F, and is computed from the formula

$$F = \frac{\text{larger variance}}{\text{smaller variance}} \qquad (9.46)$$

(F ratio for testing a difference between two estimates of a population $\tilde{\sigma}$)

Each of these estimated variances is computed by the usual method: sum of squares in the sample divided by the number of degrees of freedom. This application of the F test rests upon the assumption that the population is normally distributed.

A small set of data will illustrate the operation of this procedure. Assume that two sets of scores, in one of which $N_1 = 8$ and in the other of which $N_2 = 5$, have sums of squares $\Sigma x^2_1 = 132$ and $\Sigma x^2_2 = 26$. The degrees of freedom are 7 and 4, respectively, so the estimated variances of the population, independently derived, are 18.86 and 6.5. The F ratio is 18.86/6.5, which equals 2.90.

The Distribution of F.—In random sampling, the distribution of F ratios can be predicted from the mathematical relationships. Figure 9.8 represents three distributions for the situations with certain combinations of degrees of freedom, all of them being very small samples. Especially to be noted is the marked skewness of the curves. The test of an F ratio is made only in the tail at the right, since all the ratios examined for significance are in that region. The probability of an F's exceeding a certain value by chance is given by the area under the tail beyond that F value.

Table F (Appendix B) gives the standard F limits that are significant at the .05 and .01 levels of confidence when there are different combinations of degrees of freedom in connection with each of the two variances in the ratio. For the problem above, the two degrees of freedom are 7 and 4,

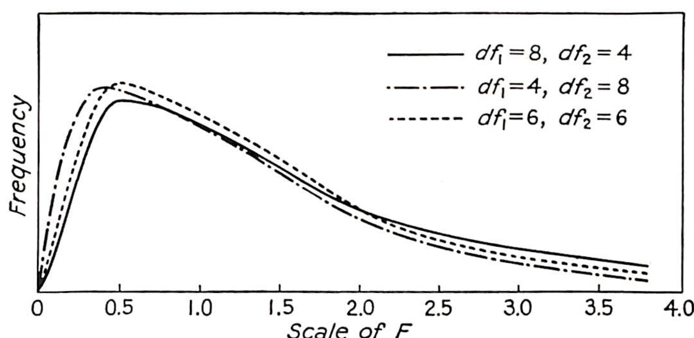


FIG. 9.8.—Sampling distribution of Snedecor's F for various combinations of degrees of freedom. (*After Lewis, D. Quantitative methods in psychology. Iowa City. The author,* 1948.)

respectively, for the larger and smaller (or numerator and denominator) variances. Looking into the appropriate column and row of Table F, we find that the two F's for the two significance levels are 6.09 and 14.98, respectively. The obtained F does not even approach the former of these very closely. We therefore do not reject the null hypothesis and decide that so far as variance or variability is concerned the two samples could well have come from the same population.

In the following chapter we shall see the F test extended considerably to the problems of analysis of variance. It is in that connection that the F test justifies the recognition that it deserves. The application demonstrated here is only one of many.

Sequential Analysis.—There has been developed very recently a procedure that enables the investigator to save considerable time and effort by testing for significance as he samples. Large differences are likely to prove significant with rather small samples. It would be wasteful of

experimental effort to accumulate more cases than would be needed to give a very significant t or F. When we have no advance information as to how large a difference is going to be, we do not know how large a sample will be needed. We could obtain a small sample, test the difference, and, if it proved significant, stop the experiment. If it did not prove significant, we would continue to add observations sampled in the same manner, then make another test, and so on. Eventually, the test goes in the direction of one hypothesis or another. This principle is applied in the method known as *sequential analysis*. There is insufficient space to describe the method adequately here. The reader is referred to an original source on the subject.[1]

Exercises

DATA 9*A*.—RESULTS FROM A TEST OF THE ABILITY TO NAME FACIAL EXPRESSIONS IN THE RUCKMICK PHOTOGRAPHS

Statistic	Men	Women
N	95	164
M	21.1	22.0
σ	3.62	3.15
Q	2.38	2.16
Mdn	21.5	22.2

DATA 9*B*.—QUANTITY WRITTEN IN SENTENCE CONSTRUCTION FROM 10 SETS OF 3 NOUNS EACH AND 10 SETS OF 3 VERBS EACH

Measurement Is the Number of Sentences Written in a Limited Time. Subjects Were 55 Girls

Statistic	Nouns	Verbs
M	24.7	22.8
σ	6.31	5.42

$r_{NV} = .67$

1. Compute the standard errors of the means for Data 9*A*, and interpret your results.

2. Compute the standard errors of the means for Data 9*B*, and interpret your results.

3. Compute the standard errors of the medians for Data 9*A*, and interpret your results.

4. Compute the standard errors of the standard deviations in either Data 9*A* or Data 9*B*, and interpret your results.

[1] Wald, A. Sequential analysis. New York: Wiley, 1947.

DATA 9C.—NUMBER OF STUDENTS IN TWO GROUPS WHO PASSED EACH OF THREE ITEMS IN AN INTRODUCTORY PSYCHOLOGY EXAMINATION

	Group I	Group II
N	37	63
Item A	24	26
		$r_{AB} = .19$
Item B	33	32
		$r_{BC} = .32$
Item C	30	44
		$r_{AC} = .25$

5. Compute the standard errors of the frequencies of passing students in Data 9C, and interpret your results. Do the same in terms of percentages and proportions.

6. Compute the standard error of the difference in means for Data 9A and also for Data 9B, and test for significance. State interpretations.

7. Compute the standard error of the difference between medians in Data 9A, and interpret your results.

8. Determine the reliability of the differences between standard deviations in Data 9A and 9B. Draw conclusions.

9. Determine the reliability of differences between Groups I and II, Data 9C, in terms of frequencies, percentages, or proportions of correct responses. Interpret your results.

10. Determine the reliability of the differences between proportions passing items A, B, and C for either Group I or Group II. Give your interpretations.

11. Assume that Data 9A are in a stratified-random sample. Compute the *SE* of the mean for a combined sample on the basis of this assumption. Compute the *SD* of such a combined sample, using formula (5.20). Compute the *SE* of the mean from this *SD* and compare it with the other. Explain the difference.

12. Assume that the same 55 girls as in Data 9B repeated the same test with the following means: 26.1 and 23.5, for nouns and verbs, respectively. The two *SD*'s were 5.12 and 5.04, respectively. The corresponding reliability coefficients (test-retest) were .87 and .75. What are the best estimates of the *SE*'s of the means in the second samples?

13. Was there a significant gain in either the noun score or the verb score? Support your answer with evidence, reporting the major steps you took.

14. The correlation between an interest score and degree of satisfaction in a certain vocational assignment was .33 in a sample of 102. Find σ_r, σ_{r_0}, and the t ratio for this finding. Interpret your results.

15. Apply Fisher's t formula for a σ_{d_m} to the following data:

$$N_1 = 11,\ N_2 = 26,\ M_1 = 17.5,\ M_2 = 14.8,\ \Sigma x^2{}_1 = 44, \text{ and } \Sigma x^2{}_2 = 65.$$

Interpret your results.

16. Test the *SD*'s in Exercise 15 for significance of their difference by making an F test. Interpret your results.

CHAPTER 10

INTRODUCTION TO ANALYSIS OF VARIANCE

It frequently happens in psychological and educational research that we obtain more than two sets of measurements, each under its own set of conditions, and we want some indication as to whether there are significant differences among the sets. We could, of course, pair off two sets at a time, pairing each one with every other one, and test the reliability of the difference in each pair. The practical difficulty in this approach lies in the number of pairs to be examined when there are, let us say, 5 or more sets. Five sets mean 10 pairs; 6 sets mean 15 pairs; and 10 sets mean 45 pairs. There is always the possibility that none of the differences would prove significant. What we desire in meeting this situation is some procedure by which we can say in advance whether or not there are *any* significant differences. If the answer to such a preliminary survey is "Yes," we can then examine pairs to see just where significance differences exist. If the answer is "No," our search is over without further ado.

The methods of R. A. Fisher, known as *analysis of variance*, are well designed to meet this kind of problem as well as other problems. The real problem here is to determine whether sets of data obtained under varying conditions are sufficiently homogeneous to be regarded as belonging to the same population. Whether or not we combine distributions into larger composite distributions sometimes hinges on the answer to this question. Fisher's test of significance in connection with his analysis of variance is designed precisely to tell us whether sets of data are sufficiently different from one another for us to reject the hypothesis that they arose by random sampling from the same population.

Analysis in a One-way Classification Problem

Total Variance in a Composite Sample.—At this time it may be profitable for the reader to review the topic of "standard deviation in a composite sample" treated in Ch. 5. At that place it was shown how the sum of squares of a composite distribution which is made up of a combination of several sets of measurements, or subsamples, is equal to the summation of sums of squares within subsamples plus the sum of squares of the subsample means around the composite mean. In terms of an equation (for-

mula 5.17), which is repeated here,

$$\Sigma x^2_t = \Sigma x^2_a + \Sigma x^2_b + n_a d^2_a + n_b d^2_b \tag{10.1}$$

This equation is limited to the combination of two subsamples, A and B. It could be extended to include any number of subsamples, with a pair of terms like those for A and B added for each additional set of data. The first two terms on the right-hand side of the equation represent sums of squares *within* the subsamples. The deviations within each subsample are from the mean of that subsample, in this case, from means M_a and M_b, respectively. The symbols d_a and d_b stand for deviations of the set means (M_a and M_b) from the composite mean (M_t). Each d^2 is multiplied by the number of cases in its set, because there are as many deviations of the size of d in each set as there are cases. They represent variations *between* samples (indirectly through variation from a common mean M_t). It is as if each deviation $X - M_t$ were made up of two components, $x_a + d_a$ in the one set and $x_b + d_b$ in the other set. The total sum of squares is likewise made up of two components; that from deviation *within* sets and that from deviations *between* sets and the common reference point, M_t. Just as we have separated the sums of squares into two distinct sources we can also separate the variances into the same sources. This is one of the fundamental concepts in analysis of variance.

Two Estimations of Population Variance.—While this illustration of the segregation of sources of variance into two components—that within sets and that between sets—is a useful basis for approaching the analysis-of-variance problem, we must hasten to take the next step. In the preceding chapter the point was repeatedly stressed that in testing significances of difference we had to think in terms of population variances rather than sample variances. We must get back to the general objective of analysis-of-variance procedures, namely, to test for significant differences between several sets of independently derived experimental samples to see whether they could or could not have arisen by random sampling from the same population.

The next key principle in this connection is that we make two distinct estimates of the population variance, one derived from the *within* sum of squares and the other from the *between* sum of squares. If these two estimates are very similar we are inclined to accept the null hypothesis, that the sets of measurements *did* arise from the same population. If the two estimates differ sufficiently, *i.e.*, to the extent that random sampling cannot reasonably account for them, we reject the null hypothesis. The testing of the two estimates of variance follows the ratio method described for the comparison of variances in the preceding chapter. The criterion is

the F-ratio test. Except in rare instances the "between" variance is greater than the "within" variance, but even when it is smaller the practice is to define F as the ratio of the between variance to the within variance.[1]

Since we want estimates of population variance, to avoid biases we divide sums of squares by the degrees of freedom rather than by N. Remembering that one degree of freedom is lost in computing each mean, let us consider how many are left for the within and the between variances. Let us assume that we have k sets of n observations each. It is very convenient, though not essential, to have the same number of observations in each set and most experiments in which analysis of variance is to be applied are designed with that in mind. Within each set, from which the *within* sum of squares is derived, there is one mean from which we lose altogether k degrees of freedom. We could write this as $k(n - 1)$ or as $N - k$. If there were 10 sets with 8 observations each, we would have $80 - 10 = 70$ degrees of freedom, or $10(8 - 1) = 70$. The between variance is estimated from the k means, which may be regarded as k independent observations. The mean of the composite is a mean of these k means and one degree is lost in this manner. This leaves $k - 1$ degrees of freedom for the between variance. For the 10 sets of 8 observations each we would have 9 degrees of freedom for the between variance. Combining the degrees of freedom for the two variances, within and between, we have 79. This checks with the number we would have if we combined all the 80 observations in one set and computed one estimate of variance.

Estimation of the Within and Between Variances.—Having determined how to find the sums of squares and also the degrees of freedom for the two estimates of population variance, we are ready for the formulas. They are

$$\text{Between variance} = \frac{\Sigma n_s d^2_s}{k - 1}$$

$$\text{Within variance} = \frac{\Sigma n_s \sigma^2_s}{k(n - 1)} = \frac{\Sigma n_s \sigma^2_s}{N - k}$$

The expression $n_s \sigma^2_s$ is equal to Σx^2_s, as was said before. We may therefore substitute Σx^2_s in the last equation. And since in most practical application of analysis of variance the sets have equal n's, we may write the two equations

[1] The temporary quotation marks for "between" variance and "within" variance here are merely for the purpose of calling attention to the fact that we are shifting meaning of those terms somewhat. Where they referred to *sample* variances before they hereafter stand for estimates of *population* variances, unless sample variance is specified. "Between" and "within" sums of squares will continue to refer to samples.

$$\text{Between variance} = \frac{n\Sigma d^2}{k - 1} \tag{10.2}$$

$$\text{Within variance} = \frac{\Sigma x^2_s}{k(n - 1)} = \frac{\Sigma x^2{}_s}{N - k} \tag{10.3}$$

The subscript may now be dropped from n, since it is a constant throughout, and also from the d's, but the subscript s is left on the x^2 to indicate that we are here dealing with the deviations from the means of the sets rather than from the grand mean of the composite.

The Solution of an Analysis-of-variance Problem.—In Table 10.1, we have four sets of observations made by the same individual on the Galton

TABLE 10.1.—WORK SHEET FOR THE ANALYSIS OF VARIANCE IN FOUR SETS OF MEASUREMENTS ON THE GALTON BAR

The Measurements (X)

	Set I	Set II	Set III	Set IV	
	114	119	112	117	
	115	120	116	117	
	111	119	116	114	
	110	116	115	112	
	112	116	112	117	
ΣX_s	562	590	571	577	2,300 ΣX
M_s	112.4	118.0	114.2	115.4	115.0 M_t
Deviations within Sets (x_s)					
	+1.6	+1.0	−2.2	+1.6	
	+2.6	+2.0	+1.8	+1.6	
	−1.4	+1.0	+1.8	−1.4	
	−2.4	−2.0	+0.8	−3.4	
	−0.4	−2.0	−2.2	+1.6	
Squares of Deviations within Sets ($x^2{}_s$)					
	2.56	1.00	4.84	2.56	
	6.76	4.00	3.24	2.56	
	1.96	1.00	3.24	1.96	
	5.76	4.00	0.64	11.56	
	0.16	4.00	4.84	2.56	
	17.20	14.00	16.80	21.20	69.20 $\Sigma x^2{}_s$
Deviations of Set Means from Grand Mean (d)					
d	−2.6	+3.0	−0.8	+0.4	
d^2	6.76	9.00	0.64	0.16	16.56 Σd^2
nd^2	33.80	45.00	3.20	0.80	82.80 $n\Sigma d^2$

bar. With a constant horizontal line of 115 mm., the subject adjusted another line to seem equal to it. The four sets were obtained under four different arrangements of conditions under which the adjustments were made. Is it likely that the observations all came by random sampling from the same general "population" of adjustments, or were there systematic differences among sets sufficient to say that the data are really not homogeneous? The following steps are followed in the solution of the type in Table 10.1:

Step 1. Compute sums and means of the sets; also the grand total ΣX and the grand mean M_t.
Step 2. For every set, compute the deviations from the set mean M_s. These are equal to $(X - M_s)$.
Step 3. Square the deviations within sets to find each x^2_s. Sum these to obtain Σx^2_s, the sum of the squares of deviations within sets.
Step 4. For each set, compute d, which equals $(M_s - M_t)$.
Step 5. Square each d, and find $n\Sigma d^2$.

With these calculations completed (see Table 10.1), we have the values we need for formulas (10.2) and (10.3). The Σx^2_s is 69.20, and the $n\Sigma d^2$ is 82.80. Dividing these by the appropriate degrees of freedom, we obtain the variances. For this purpose, we set up Table 10.2. Listing first the degrees of freedom and sums of squared deviations for "between

TABLE 10.2.—THE TOTAL VARIANCE IN THE GALTON-BAR DATA SUBDIVIDED INTO TWO COMPONENTS

Components	Degrees of freedom	Sums of squares	Variance
Between sets............	3	82.80	27.60
Within sets.............	16	69.20	4.325
Total................	19	152.00	

$$F = \frac{27.6}{4.325} = 6.38$$

sets" and dividing, we obtain 27.60 as the variance contributed by the d's. For the corresponding values for "within sets," we find 4.325 as the variance contributed by the x_s's. The F ratio is 27.6/4.325, which equals 6.38. The between variance is over 6 times as great as the within variance.

The significance of an F ratio of this size is determined by reference to Snedecor's table (Table F, Appendix B). In using this table, we have to consider the two different degrees of freedom. For the larger variance,

with 3 degrees of freedom, we look for the column in Table F that is headed (3). For the smaller variance, with 16 degrees of freedom, we look down the left-hand margin for the row headed (16). We must interpolate, since row (16) is not given, and thus we find that an F of 3.24 is significant at the 5 per cent level and an F of 5.29 is significant at the 1 per cent level; *i.e.*, the odds are 5 to 95 that so large an F as 3.24 could have occurred in a really homogeneous population, and they are 1 to 99 that an F as large as 5.29 could have occurred likewise. Our obtained F is greater than that for the 1 per cent level and so is regarded as very significant. We conclude that there are significant differences among our sets. The test does not tell us where those differences are or whether all of them or only one is significant. To determine this would require further search. We only know from the F test that some significant law of variation between sets does exist. Further examination is needed to tell us what the causes of difference are and where they lie.

Making t Tests Following an F Test.—Suppose that F turns out to be significant and we want to know just where the significance in the data exists. Particularly if F is significant between the 5 per cent and 1 per cent levels, it is not likely that all of the means are significantly different from all other means. Even when F is significant at the 1 per cent level, if one mean stands out as very different from the rest and the others differ very little, it is not likely that all mutual differences are significant. We are often inclined, then, to test at least some of the differences between pairs of means. The best procedure of this is to apply Fisher's formula (formula 9.39), for a t test in small samples. We assume the null hypothesis for each pair in turn as we test it. We could save ourselves unnecessary work by being judicious in starting the t tests. For example, if F is just barely significant at the 5 per cent level, we might begin with the largest difference and proceed with other differences until one proves insignificant after which we forego further testing on the assumption that we have reached the probable limit. This would be safe, particularly if the sets have similar dispersions. If the F ratio is decidedly significant beyond the 1 per cent level, we might begin at the other end, with the smallest differences, and work up to the difference of a size that proved significant by t test, assuming that all differences as large or larger are also significant.

Some writers recommend that in making t tests after an F test we make only one estimate of population variance for all pairs and that this estimate be the within variance used in making the F test. This hardly seems logical, for if the F test has already told us that we may *not* assume that the sets all could have arisen by random sampling from the same population, it is inconsistent to make *one* estimate as if it *were* one population. The

procedure described above seems preferable, though requiring more effort.

The Relation of t to F.—When we are reduced to two sets of observations, as when we compare two means for significance, we can still make an F test. The between variance will have associated with it only 1 degree of freedom, and when this is the case it has been shown that F is equal to t^2. For this particular situation, when $N_1 = N_2$, the sum of squares for between-means variations is given by the formula

$$n\Sigma d^2 = \frac{n(M_1 - M_2)^2}{2} \quad \text{(Sum of squares between means of two samples of equal size)} \qquad (10.4)$$

To illustrate, let us take the largest difference between means in Table 10.1. The two means are 112.4 and 118.0, and their difference is 5.6. Applying formula (10.4) we find 78.4 for the between sum of squares. The within sum of squares is a combination of 17.2 and 14.0 from Table 10.1. With 1 degree of freedom for the between sum of squares, the between variance is 78.4. With 8 degrees of freedom within the sets, the within variance is $31.2/8 = 3.9$. The F ratio is $78.4/3.9 = 20.10$, which is well beyond that required for significance at the 1 per cent level, the latter being 11.26. It is an established fact that with 1 degree of freedom for the between variance, F equals t^2. If F is equal to t^2, t in this problem must therefore be equal to 4.48.

Let us check this by computing t in the usual manner, using formula (9.40). By that approach,

$$t = \frac{5.6}{\sqrt{\frac{17.2 + 14.0}{5(5 - 1)}}} = \frac{5.6}{1.25} = 4.48$$

It can be demonstrated mathematically that $t^2 = F$ under these conditions, if one starts by squaring both sides of formula (9.40). Comparison of Table D, last column (t values) and Table F, first column of F values, will show that for the same number of degrees of freedom within the sets $F = t^2$.

Computation of Variances from Original Measurements.—Just as we can compute standard deviations, and so variances, from original measurements without computing separate deviations from the means, (see formula 5.11) so we can calculate the necessary constants for an analysis of variance. Such an approach requires us to square the original measure-

ments. With good calculating machines available, this is no large order, but with only pencil and paper it amounts to considerable labor.

Fortunately, by a process of coding, we can bring the numbers down to small size. From each of the three-place numbers in Table 10.1, we may substract the constant of 110, leaving the remainders shown in the first part of Table 10.3. The variances will not have been affected in the least

TABLE 10.3.—SOLUTION OF AN ANALYSIS OF VARIANCE FROM ORIGINAL MEASUREMENTS (Without Determining Deviations from Means)

Measurements (Reduced) (X')

	Set I	Set II	Set III	Set IV		
	4	9	2	7		
	5	10	6	7		
	1	9	6	4		
	0	6	5	2		
	2	6	2	7		
$(\Sigma X')_s$	12	40	21	27	100	$\Sigma X'$
					5.0	M'_t
$(\Sigma X')^2_s$	144	1,600	441	729	2,914	$\Sigma(\Sigma X')^2_s$

Squared Measurements (X'^2)

	Set I	Set II	Set III	Set IV		
	16	81	4	49		
	25	100	36	49		
	1	81	36	16		
	0	36	25	4		
	4	36	4	49		
$(\Sigma X'^2)_s$	46	334	105	167	652	$\Sigma(\Sigma X'^2)_s$

by this particular coding process, for the new values, which we shall call X', maintain the same distances from one another and from the means as they did before coding. The sums of squares we need for equations (10.2) and (10.3) are found by the following procedure. The sum of the between variations squared is given by

$$n \sum d'^2 = \frac{\Sigma(\Sigma X')^2_s}{n} - \left(\sum X'\right)(M'_t) \tag{10.5}$$

The within sum of squares is given by

$$\sum x^2_s = \sum\left(\sum X'^2\right)_s - \frac{\Sigma(\Sigma X')^2_s}{n} \tag{10.6}$$

The total sum of squares is given by

$$\Sigma x^2 = \Sigma(\Sigma X'^2)_s - (\Sigma X')(M'_t) \tag{10.7}$$

The steps called for by these formulas are as follows:

Step 1. Sum the coded measurements X' for each set, to obtain $(\Sigma X')_s$ for each set (see Table 10.3), and sum these values to obtain $\Sigma X'$. Determine the mean M'_t to two or more decimal places.
Step 2. Square the sums of the scores to obtain $(\Sigma X')^2_s$ for each set. Accumulate these to find $\Sigma(\Sigma X')^2$.
Step 3. Square all the coded measurements to find the X'^2 values.
Step 4. Sum all the squared measurements to obtain $\Sigma(\Sigma X')^2_s$.

Now, by formula (10.5),

$$n \sum d'^2 = \frac{2{,}914}{5} - 500 = 582.8 - 500 = 82.8$$

By formula (10.6),

$$\sum x^2_s = 652 - \frac{2{,}914}{5} = 652 - 582.8 = 69.2$$

And by formula (10.7),

$$\Sigma x^2 = 652 - (100)(5) = 652 - 500 = 152$$

A check for accuracy of computation is to see that $n\Sigma d^2 + \Sigma x^2_s = \Sigma x^2$. The check is satisfied here for $82.8 + 69.2 = 152.0$.

A comparison of these values with those in Table (10.2) will show that we have arrived at the very same sums of squares. From here on the computation of variances and of F ratio is just the same as it was before. The same formulas, (10.5) through (10.7), apply also to original measurements without coding.

Analysis in a Two-way Classification Problem

In the preceding problem there was only a one-way classification. The sets of data were differentiated on the basis of only one experimental variation, or were at least treated as if there was only one reason for fractionating the data into sets. The principle of division into sets might have been time of day, degree of learning, of fatigue, or of illumination. The variable, if it was a single one, in which there were differences from set to set might have been a quantitative one, as in the examples last mentioned, or

a qualitative one. An example of a qualitative basis of classification in a maze-learning experiment would be several learning methods, such as verbal instruction, demonstration, "putting through," and prevention of errors. A quantitative basis, also in maze learning, might be different lengths of time for visual inspection of the maze before starting to learn it.

In a two-way classification, there are two distinct bases of classification. Two experimental conditions are allowed to vary from trial to trial. Usually there are several trials made under each *combination* of conditions. In the psychological laboratory a study of different air-field landing strips each with a different pattern of markings may be viewed through a diffusion screen to stimulate vision through fog at different levels of opaqueness. In an educational problem, four methods of teaching a certain geometric concept may be applied by five different teachers, each one applying every one of the four methods. There would therefore be 20 combinations of teacher and method, and let us suppose an equal number of pupils each giving a learning score under each combination.

Tabulation of Data in a Two-way Classification Problem.—For an illustration of the procedure here, we will assume an experiment on the relation of scores on a certain psychomotor test to the size of a target at which the examinee must aim. In conducting the experiment it is convenient to use three testing machines simultaneously in order to reduce the testing time. It is known that there are individual differences between machines, in this test, to the extent that it would be risky to attach one target size to one machine only throughout the tests. Machine differences might make it appear that there were differences attributable to target differences or might by chance negate those differences. The target sizes were therefore combined with the machines systematically. There were therefore 12 target-machine combinations with 5 observed scores obtained with each combination. The scores (which are entirely fictitious for the sake of a good illustration) are tabulated in Table 10.4. This arrangement is typical and convenient for the operations of analysis of variance. The sums and means, as given, are also needed in the variance solution.

The Sources of Variance in a Two-way Classification Problem.—We could, if we chose, proceed to perform an analysis of variance based upon the model of the one-way classification problem as already demonstrated. That is, we could take the 12 sets as if they represented categories based upon a single principle and test the 12 means collectively to see whether they could have arisen by random sampling from the same population. We shall see what kind of an answer could be obtained by this approach, later, but let us first see what is logically wrong with this kind of solution here.

TABLE 10.4.—SCORES OF 60 STUDENTS EARNED ON THREE DIFFERENT MACHINES OF A PSYCHOMOTOR TEST, EACH WITH THE TARGET SIZE VARIED IN FOUR STEPS

Target size		Machines			Sums for target size	Means for target size
		1	2	3		
A		6	4	4		
		4	1	2		
		2	5	2		
		6	2	1		
		2	3	1		
	Σ	20	15	10	45	
	M	4	3	2		3
B		8	6	3		
		3	6	1		
		7	2	1		
		5	3	2		
		2	8	3		
	Σ	25	25	10	60	
	M	5	5	2		4
C		7	9	6		
		6	4	4		
		9	8	3		
		8	4	8		
		5	5	4		
	Σ	35	30	25	90	
	M	7	6	5		6
D		9	7	6		
		6	8	5		
		8	4	7		
		8	7	9		
		9	4	8		
	Σ	40	30	35	105	
	M	8	6	7		7
Sums for machines. .		120	100	80	300	
Means for machines.		6	5	4		5

Suppose we did carry through the solution proposed and found an *F* ratio that indicated significance beyond the 1 per cent level. We would not know whether this was due primarily or solely to the differences between targets or to the differences between machines, or to both possible sources. Suppose, on the contrary, the *F* ratio indicated no significant differences among sets. We would not be sure that one of the experimental variations, perhaps target size, were not actually producing real

variations that were either covered over or counteracted by the effects of the other experimental variation. We need some method that will segregate the variations associated with each of the experimental variables so that any significant differences at all will have a chance to emerge in the F test and so that we will know to which source to attribute any significant differences found.

Interaction Variance.—The procedure about to be described makes possible this kind of segregation of the sources of variations. As a result, we can then determine whether differences among means owe their divergencies to target size or to machine differences, or to both. Not only that, when there are two possible sources of variations, there is also a possibility of what is called *interaction variance.* The phenomenon is well named. Interaction variations are those attributable not to either of two influences acting alone but to joint effects of the two acting together. If it turned out that the larger the target the larger the scores tended to be, that is one direct and isolable effect. If there are systematic machine differences so that among three there is a most "difficult" one (yields lower mean scores) and an easiest one (yields higher mean scores), that is another distinct effect. There may be effects of target size and machine over and above these. It is conceivable, but not very probable, that one machine, apart from its general difficulty, gains in difficulty by virtue of its having one size of target rather than others. It may be the coincidence of machine and target size that produces systematic variation in one direction from the general mean of scores. This is an example of interaction variance. It might be more reasonably expected in combination of teacher and instruction method; of kind of task and method of attack by the learner; and of kind of reward when combined with a certain condition of motivation. It is also possible to determine whether there is a significant amount of interaction variance present by making an F test for it.

The Residual Variance.—There are three F tests to make, therefore, in place of one. The remaining variance is known as the residual variance, that within sets. It supplies the basic or residual estimate of variance after the three sources of variations have been removed and it serves as the denominator for all three F tests. It is sometimes called an estimate of the *error* variance for the reason that it represents the influences of many unknown and uncontrolled sources. A perfect experiment would presumably control all contributing factors until within each set of data observed under a specified combination of conditions there would be no longer any variations; each observed value would be the same. Most experiments are so imperfect that there is appreciable error variance.

Estimation of the Variances from Different Sources.—Two solutions will be described, one using deviations of observed values and of means of sets from various means, the other using original measurements and means. An attempt is made to summarize the operations in terms of formulas, as usual, but here the symbolizing of concepts becomes so involved that formulas may be more confusing than helpful. Some readers may find it easier to follow the examples as models rather than to apply the formulas. The systems of symbols employed in the formulas is given in Table 10.5. This table provides only three columns and three rows, but it can be extended in the directions shown to take care of any number of columns and rows.

TABLE 10.5.—SYMBOLIC SCHEME FOR THE VALUES IN A TABULATION PREPARATORY TO ANALYSIS OF VARIANCE IN A TWO-WAY CLASSIFICATION PROBLEM

Row		Column			Sums of rows (ΣX_r)	Means of rows (M_r)
		1	2	3		
	1	X_{a1}	X_{a2}	X_{a3}		
	2					
	3					
A	4					
	5					
	Σ	ΣX_{a1}	ΣX_{a2}	ΣX_{a3}	ΣX_a	
	M	M_{a1}	M_{a2}	M_{a3}		M_a
	1	X_{b1}	X_{b2}	X_{b3}		
	2					
	3					
B	4					
	5					
	Σ	ΣX_{b1}	ΣX_{b2}	ΣX_{b3}	ΣX_b	
	M	M_{b1}	M_{b2}	M_{b3}		M_b
	1	X_{c1}	X_{c2}	X_{c3}		
	2					
	3					
C	4					
	5					
	Σ	ΣX_{c1}	ΣX_{c2}	ΣX_{c3}	ΣX_c	
	M	M_{c1}	M_{c2}	M_{c3}		M_c
Sums of columns (ΣX_k).		ΣX_1	ΣX_2	ΣX_3	ΣX_{ij}	
Means of columns (M_k).		M_1	M_2	M_3		M_t

Let X_{ij} = any one of the cell entries, X_{a1}, X_{b1}, . . . X_{c3}.
M_{rk} = any one of the set means, M_{a1}, M_{a2}, . . . M_{c3}.

The Solution Based upon Deviations.—In what follows, consistent with the symbols in Table (10.5), a subscript k stands for a particular column (we might have used c for column, but there would be danger of confusing this with a particular row—row C), and r stands for a particular row. There are only three columns, 1, 2, and 3, in the psychomotor test problem, and four rows, A, B, C, and D. The symbol X_{ij} stands for any one observation in row r and column k and M_{rk} stands for a mean of the five observations in a cell described as being in row r and column k. In the following, n stands for the number of observations within each set; in the illustrative problem $n = 5$. The number of rows is symbolized by r and the number of columns by k. The subscript t refers to the total distribution, all sets combined. Thus, M_t stands for the mean of the composite, and x_t stands for a deviation of any X from M_t.

The total sum of squares is given by the equation

$$\Sigma x^2_t = \Sigma(X_{ij} - M_t)^2 \qquad (10.8)$$

Applied to the data of Table 10.4,

$$\begin{aligned}
\Sigma x^2_t &= (6 - 5)^2 + (4 - 5)^2 + (4 - 5)^2 \text{ (from first row of Table 10.4)} \\
&\quad + \cdots\cdots\cdots \\
&\quad + (9 - 5)^2 + (4 - 5)^2 + (8 - 5)^2 \\
&\qquad\qquad \text{(from last row of observations in Table 10.4)} \\
&= 1^2 + (-1)^2 + (-1)^2 \\
&\quad + \cdots\cdots \\
&\quad + 4^2 + (-1)^2 + 3^2 \\
&= 374 \text{ (total sum of squares)}
\end{aligned}$$

The sum of squares between rows is given by the equation

$$\Sigma d^2_r = nk[\Sigma(M_r - M_t)^2] \qquad (10.9)$$

Applied to the same data,

$$\begin{aligned}
\Sigma d^2_r &= 5 \times 3[(3 - 5)^2 + (4 - 5)^2 + (6 - 5)^2 + (7 - 5)^2] \\
&= 15[(-2)^2 + (-1)^2 + 1^2 + 2^2] \\
&= 15 \times 10 \\
&= 150 \text{ (the sum of squares between rows)}
\end{aligned}$$

The sum of squares between columns is given by the equation

$$\Sigma d^2_k = nr[\Sigma(M_k - M_t)^2] \qquad (10.10)$$

Applied to the data of Table 10.4,

$$\begin{aligned}\Sigma d^2_k &= 5 \times 4[(6 - 5)^2 + (5 - 5)^2 + (4 - 5)^2] \\ &= 20[1^2 + (-1)^2] \\ &= 20 \times 2 \\ &= 40 \text{ (sum of squares between columns)}\end{aligned}$$

The interaction variance can be estimated in several ways. Perhaps the most common way is to derive it from the sum of squares between all sets, eliminating the sums of squares between columns and between rows. We already know the last two sums of squares. We proceed next to compute the sum of squares between sets. The formula is similar to the numerator of formula (10.2) but with different notation to fit the new system.

$$\Sigma d^2_{rk} = n[\Sigma(M_{rk} - M_t)^2] \qquad (10.11)$$

The symbol d^2_{rk} refers to a squared difference between any set mean and the total mean M_t. The subscript rk implies that all rows and all columns are involved. Applied to the illustrative data,

$$\begin{aligned}\Sigma d^2_{rk} &= 5[(4 - 5)^2 + (3 - 5)^2 + (2 - 5)^2 \text{ (from first row of means)} \\ &\quad + \cdots\cdots\cdots\cdots \\ &\quad + (8 - 5)^2 + (6 - 5)^2 + (7 - 5)^2] \text{ (from last row of means)} \\ &= 5[(-1)^2 + (-2)^2 + (-3)^2 \\ &\quad + \cdots\cdots\cdots \\ &\quad + 3^2 + 1^2 + 2^2] \\ &= 5 \times 42 \\ &= 210 \text{ (sum of squares between means of sets)}\end{aligned}$$

If we remove from entire sum of squares for the 12 set means the sum of squares attributable to columns and to rows, we have left the interaction sum of squares. By formula,

$$\Sigma d^2_{r \times k} = \Sigma d^2_{rk} - \Sigma d^2_k - \Sigma d^2_r \qquad (10.12)$$

in which $\Sigma d^2_{r \times k}$ (the subscript reads r times k, for reasons that will be explained) stands for the interaction sum of squares. For the illustrative problem,

$$\begin{aligned}\Sigma d^2_{r \times k} &= 210 - 40 - 150 \\ &= 20 \text{ (interaction sum of squares)}\end{aligned}$$

Another, more direct, way of deriving interaction sums of squares utilizes the formula

$$\Sigma d^2_{r \times k} = n\,[\Sigma(M_{rk} - M_k - M_r + M_t)^2] \qquad (10.13)$$

in which M_k is the mean of the column in which each particular M_{rk} appears and M_r the mean of its row. For the illustrative problem,

$$\begin{aligned}\Sigma d^2_{r\times k} &= 5[(4-3-6+5)^2+(3-3-5+5)^2 \text{ (from first row of means)}\\ &+ \cdots\cdots\cdots\cdots\\ &+ (6-7-5+5)^2+(7-7-4+5)^2] \text{ (from last row of means)}\\ &= 5[(0^2+0^2+\cdots+(-1)^2+1^2]\\ &= 5\times 4\\ &= 20 \text{ (interaction sum of squares; alternative solution)}\end{aligned}$$

The sum of squares within sets is computed by the formula

$$\Sigma x^2_s = \Sigma(X_{ij} - M_{rk})^2 \tag{10.14}$$

This formula, with new symbols, requires the same operations as formula (10.3) given in connection with the single-classification problem. Applied to the psychomotor problem,

$$\begin{aligned}\Sigma x^2_s &= (6-4)^2+(4-4)^2+(2-4)^2+(6-4)^2+(2-4)^2\\ &\qquad\qquad \text{(from set } A1)\\ &+ \cdots\cdots\cdots\cdots\\ &+ (6-7)^2+(5-7)^2+(7-7)^2+(9-7)^2+(8-7)^2\\ &\qquad\qquad \text{(from set } D3)\\ &= 164 \text{ (sum of squares within sets)}\end{aligned}$$

We can now check the solution of this by deducting all previously computed sums of squares from the total sum of squares, and we have

$$374 - 40 - 150 - 20 = 164$$

We could compute Σx^2_s by this elimination process without going through the arduous arithmetic involved in using (10.14), but for checking purposes it is very desirable to derive all the component sums of squares separately and then check the results.

Degrees of Freedom.—Before taking the next important step of estimating population variance from these different sources, we need, as usual, the degrees of freedom. Starting with the largest source, the total sum of squares, we have, as usual, $(N - 1)$ *df*, or 59. This figure is to be subdivided among the contributing components. The sum of squares among the means of sets should have allotted to it the number of sets minus 1, or $12 - 1 = 11df$. These 11, in turn, are to be allotted to three sources. Rows have the number of row observations (row means) minus 1, or $4 - 1 = 3$. Columns have, by analogy, $3 - 1 = 2$. This leaves 6 out of the 11 for interaction. This 6 degrees is 3×2, the product of the *df* for rows and columns each source taken separately. This is consistent with

the idea of interaction itself, whose contributions to variations may be regarded as the products of two sources. This is why we use the subscript $r \times k$ when referring to interaction. Having taken care of the special sources of variations, the remainder, or 59 − 11, gives us the *df* left for within-sets sums of squares. This number of *df* may also be determined directly from a summation of *df* within sets. Since there are 12 sets and each contains 4 *df*, we have $12 \times 4 = 48$ *df* for the residual variance.

In terms of symbolic descriptions, the degrees of freedom may be given as follows:

Source	*Degrees of Freedom*
Between rows	$r - 1$
Between columns	$k - 1$
Interaction	$(r - 1)(k - 1)$
Within sets	$N - rk = rk(n - 1)$
Total	$N - 1$

The *F* Ratios.—We are now ready to estimate the variances and to compute the *F* ratios. These are systematically arranged in Table 10.6. There are four different estimates of population variance—50.0, 20.0, 3.33, and 3.42. We compare the first three, since they represent possible special contributions resulting from varied experimental conditions, each with the fourth. The fourth presumably represents variations of the phe-

TABLE 10.6.—SOURCES OF VARIANCE IN THE PSYCHOMOTOR-TEST DATA ANALYSIS AND *F* RATIOS

Source	Sum of squares	Degrees of freedom	Estimate of variance
Target size (T)	150	3	50.0
Machine (M)	40	2	20.0
Interaction ($T \times M$)	20	6	3.33
Within sets	164	48	3.42
Total	374	59	

	Required *F* 5% level	Required *F* 1% level
F for targets $= \frac{50}{3.42} = 14.62$	2.80	4.22
F for machines $= \frac{20}{3.42} = 5.85$	3.10	5.08
F for interaction $= \frac{3.33}{3.42} = 0.97$	2.30	3.20

nomenon measured freed from possible influences of the experimental variations. Do the first three differ significantly from the fourth?

The F ratios are given below the table, together with the F's required for significance at the 5 per cent and 1 per cent levels as determined from Snedecor's table (Table F). From these results it appears that variations in target size definitely carry with them systematic variations in test score. There is a law of relationship fairly well established between target size and difficulty of the test. The F ratio for machines is significant beyond the 1 per cent level, leaving us with considerable confidence that the machine differences, as such, have a real bearing upon the difficulty of the task. This conclusion is in some doubt because of possible failure of experimental design, however. Since the examinees were different groups for the three test machines, we cannot be sure that some real differences of ability have not combined with minor machine differences to give an apparently significant machine difference. A matching of examinees for machines might have improved the precision of the experiment. This would have entailed modification in the analysis-of-variance operations. The F for interaction proved to be rather decidedly insignificant. There is no reason to believe that changing target size has different effects depending upon the machine with which it is associated.

Removal of Sources of Variation.—It may illuminate the concepts of different kinds of variance and the way in which they contribute to total variance in the sample if we separate them in another way. Table 10.7*A* shows the 12 means of sets for the psychomotor-test data. Variations among them are due to the three possible sources—target differences, machine differences, and the interaction of the two. The possible effects of target size are most apparent in the means of the rows—3, 4, 6, and 7. The possible effects of machine differences are most apparent in the means of the columns—6, 5, and 4. The possible interaction variance is obscured. It possibly contributes both to the means of rows and of columns; we do not know. Let us strip away first the variations attributable to machines and then that attributable to targets and see what variations are left.

The mean of all observations is 5. Any deviation of a column mean from 5 indicates a constant error for a particular machine. Machine 1 gave a mean of 6, indicating that machine 1 had a constant error of +1. Machine 2 apparently had no constant error, while machine 3 had a constant error of −1. If we deduct from each cell or set mean in column (1) the amount of constant error involved for machine 1, we would persumably remove from the means in column (1) the influence of machine 1 as a source of variation. We can do likewise for column (3), deducting the constant error of −1, which is equivalent to adding +1 to each mean.

TABLE 10.7.—ANALYSIS OF THE BETWEEN-SETS SUMS OF SQUARES IN THE PSYCHOMOTOR-TEST DATA INTO THREE COMPONENTS BY SUCCESSIVE REMOVAL OF CONTRIBUTING SOURCES OF VARIATION

Row	Column			Σ	*M*
	1	2	3		
A. Original Matrix of Means of Sets					
A	4	3	2	9	3
B	5	5	2	12	4
C	7	6	5	18	6
D	8	6	7	21	7
Σ	24	20	16	60	
M	6	5	4		5
B. With Variations Associated with Machines Removed					
A	3	3	3	9	3
B	4	5	3	12	4
C	6	6	6	18	6
D	7	6	8	21	7
Σ	20	20	20	60	
M	5	5	5		5
C. With Variations Associated with Target Size Also Removed; Only Interaction Variance Remaining					
A	5	5	5	15	5
B	5	6	4	15	5
C	5	5	5	15	5
D	5	4	6	15	5
Σ	20	20	20	60	
M	5	5	5	..	5

We need do nothing for column (2). The results of these operations are shown in Table 10.7*B*. The means of the columns are now all 5, to agree with the composite mean, M_t. The means of the rows have been unaffected (they are still 3, 4, 6, and 7) because the changes in one column are compensated for by changes in reverse direction in another column. The cell values in Table 10.7*B* still have in them the variance attributable to targets and to interaction variance.

Next we remove the target variance. The constant errors for rows are -2, -1, 1, and 2, respectively. Deducting these from the values in their

respective rows in Table 10.7*B*, we have the results in subtable *C*. The means of the rows as well as of the columns are now all 5. But within four cells there are departures from 5. These are possibly the interaction deviations, depending upon whether or not they prove to be significant. Machine 2 would seem to favor high scores when coupled with target *B* and to favor low scores when coupled with target *D*. Machine 3 has a reverse tendency. But the *F* test showed these deviations to be insignificant. There seem to be no good logical reasons to expect any systematic coupling of target and machine. In other problems there may be significant interaction effects, but one would expect them to be systematic when experimental variables are quantitative in character.

The finding of insignificant deviations among the means suggests several things. One is that these variations are random sampling effects that really belong to the within variance but were not pulled out with it. There is good reason, therefore, for combining this source of variance with that from within sets. The sum of squares for this was 20. Combined with that from within sets, we have a total of 184. With 48 degrees of freedom, we have a within variance that is raised from 3.42 to 3.83. This change is not enough to make any material difference in the *F* ratios for the target or machine sources. Our inferences about those sources being significant remain unchanged.

Second Solution—From Original Measurements.—Next will be given the formulas and their application for the solution of means of squares without computing deviations. With small integral numbers to start with, or numbers coded to such magnitude, these procedures are often more convenient than those utilizing deviations. The first solution, with deviations, is more meaningful to the beginner. In the following exposition, each formula will be stated then immediately applied to the psychomotor-test data.

Total sum of squares:

$$\sum x^2_t = \sum X^2_{ij} - \frac{(\Sigma X_{ij})^2}{N} \tag{10.15}$$

$$= (6^2 + 4^2 + 4^2 \text{ (from first row of Table 10.4)}$$
$$+ \cdot\cdot\cdot\cdot\cdot\cdot$$
$$+ 9^2 + 4^2 + 8^2)$$
$$\text{(from last row of observations in Table 10.4)}$$
$$- \frac{(300)^2}{60}$$
$$= 1874 - 1500$$
$$= 374 \text{ (total sum of squares)}$$

Sum of squares between sets:

$$\sum d^2_{rk} = \frac{\Sigma(\Sigma X_{rk})^2}{n} - \frac{(\Sigma X_{ij})^2}{N} \tag{10.16}$$

$= \frac{1}{5}\,[(20^2 + 15^2 + 10^2$(from first Σ row of Table 10.4)

$+ \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot$

$+ 40^2 + 30^2 + 35^2)]$ (from last Σ row of Table 10.4)

$- \frac{(300)^2}{60}$

$= 1710 - 1500$

$= 210$ (sum of squares between sets)

Sum of squares between rows:

$$\sum d^2_r = \frac{\Sigma(\Sigma X_r)^2}{nk} - \frac{(\Sigma X_{ij})^2}{N} \tag{10.17}$$

$= [\frac{1}{15}(45^2 + 60^2 + 90^2 + 150^2)] - 1500$

$= 1650 - 1500$

$= 150$ (sum of squares between rows)

Sum of squares between columns:

$$\sum d^2_k = \frac{\Sigma(\Sigma X_k)^2}{nr} - \frac{(\Sigma X_{ij})^2}{N} \tag{10.18}$$

$= [\frac{1}{20}(120^2 + 100^2 + 80^2)] - 1500$

$= 1540 - 1500$

$= 40$ (sums of squares between columns)

Sum of squares for interaction:

$$\Sigma d^2_{r \times k} = \Sigma d^2_{rk} - \Sigma d^2_r - \Sigma d^2_k \tag{10.19}$$

$= 210 - 150 - 40$

$= 20$ (sum of squares for interaction)

Sum of squares within sets:

$$\Sigma x^2_s = \Sigma x^2_t - \Sigma d^2_{rk} \tag{10.20}$$

$= 374 - 210$

$= 164$ (sum of squares within sets)

It will be noted that the correction factor $(\Sigma X_{ij})^2/N$, which appears in most of these equations, is identical and once computed will do thereafter.

The sums of squares by this method are seen to be identical with those found by the preceding method. The estimation of the population variance from each source and the application of the F test are the same as before (see Table 10.6).

An Evaluation of Analysis of Variance

Assumptions to Be Satisfied in Applying Analysis of Variance.—Like most statistics, those involved in analysis of variance have been derived on the basis of mathematical reasoning, and that reasoning starts with certain assumptions. If those assumptions are satisfied within certain limits of tolerance, the results in terms of F ratios may be interpreted as described in this chapter. If those assumptions are not sufficiently approximated, there is considerable risk that the conclusions may be faulty.

There are four assumptions often specified. They are:[1]

1. The contributions to variance in the total sample must be additive. This is implied in the equation (10.1), in which the total sum of squares is assumed to be a summation of sums of squares from within sets plus sums of squares from between sets. The same summative idea is illustrated in Table 10.7 in which we stripped off one by one the three sources of variance. The additive nature of variations squared is dependent to some extent upon other assumptions to follow.

2. The observations within sets must be mutually independent. The "laws of chance" must be allowed to operate in an unrestricted way. The occurrence of a certain deviation in one observation must be in no way dependent upon any other deviation. This is, of course, a good description of random sampling. The random sampling occurs within sets. The intentional variation of experimental conditions may produce systematic variations between sets. Whether or not systematic variations do occur is the thing being tested.

3. The variances within experimentally homogeneous sets must be approximately equal. By "experimentally homogeneous" is meant observations under one specified set of experimental conditions. The "within-set" variance, is, of course, the denominator of every F ratio. It therefore carries a heavy burden, especially if there is more than one F to be computed for the same data. This variance is used as a single estimate of the population variance, and all contributors to it should tell the same story. If any two sets furnish widely divergent ideas of the population variance, the latter is not very accurately estimated. If there is serious doubt about the variances indicated by any two sets, we can, and should, make an F test for the difference between those two variances. If F is so high as to cause rejection of the null hypothesis, we should not use both set results as sources of a within sum of squares.

[1] Cochran, W. G. Some consequences when the assumptions for the analysis of variance are not satisfied. *Biometrics*, 1947, **3**, 22–38.

4. The variations within experimentally homogeneous sets should be normally distributed. Most subsamples will be so small that it will be out of the question to test the distribution for normality by the chi-square test. Decision as to normality or lack of it will have to come from other sources. Knowledge of the form of population distribution would be sufficient basis, if sampling is random. Skewness, if marked, is a serious source of violation of the validity of an F test.

If we follow the practice of free and random sampling within sets and if we use a metric scale on which there is lack of restriction and on which units are equal, we can feel assurance that the F test will not be invalidated. It must be remembered, however, that conditions of sampling are never ideal. F tests are therefore usually only approximate. Under somewhat doubtful circumstances, an F that proves to be significant at the 5 per cent level may be actually significant anywhere from the 4 to the 7 per cent level; one significant at the 1 per cent level might actually be significant between the 0.5 per cent and 2 per cent level.[1] If anything, the significance is likely to be *lower* than that indicated by the result, when assumptions are not well satisfied.

General Uses and Limitations of Analysis of Variance.—There is insufficient space here to do more than to give this introduction to the analysis-of-variance methods. There are many and varied applications of these basic cases—the separation of variance among a few sets of data into the "within" and "between" variances—both in psychology and in education.

Sets of data may be divided according to chronological-age groups, mental-age groups, sex-difference groups, etc. In psychophysical experiments, judgments of phenomena may be made under various conditions—ascending series versus descending series, variable stimulus first versus second, right versus left, and with many other kinds of variation of conditions, not to speak of individual differences among observers, and observations obtained at different times of day or under different states of fatigue or rest or after different degrees of practice. In education, the testing of different teaching methods can be done in different schools, in different classes within the same school, and with different teachers.

It will be recognized that the conditions affecting sets of measurements often vary in several directions at the same time. This complicates the analysis-of-variance solutions in various ways. There are also problems in which the sets of data are not independently observed, as was assumed in the present chapter. There is a technique for the analysis of covariance as well as of variance. Covariance and correlation are closely related, as

[1] Cochran, *op. cit.*

will be shown in later chapters. For further descriptions of how to adapt the method to various kinds of experimental problems, the reader is referred to books that treat the subject at much greater length.[1]

By way of hasty evaluation of the method, it may be said that analysis of variance undoubtedly provides a powerful tool of working through data in order to see where the significant lines of cleavage lie and thus furnishes some basis for establishing the presence of laws. It can also be said that the method requires supplementary procedures for a more detailed study of data and that there are other statistical methods—for example, correlation procedures—that enable us to accomplish the same purpose in many instances.

Not the least of its merits is the rather strict set of requirements it presupposes in the designing of experiments. Experimental designs have generally been observed, particularly in psychophysical research, for a long time. But they have generally not been so consciously considered or so well planned so as to yield the maximum number of dependable answers as is true when the experimenter has kept clearly in mind the corresponding statistical tests that go with those designs. The subject of experimental design is well treated in the book by Lindquist already cited, so far as certain educational problems are concerned. Discussions of designs for psychological and other experiments may be found elsewhere.[2]

Exercises

Assume that Data 10*A* represent measurements of the lower threshold for pitch of tones under the following conditions. The observer was the same throughout. Each "trial" was composed of 100 observations, of which four were selected at random. The 400 observations were made during the same half day, with only short rest pauses after each 25 and with 10 minutes between trials.

1. Using the four sets of observations made on the first day, apply an F test to determine whether there were systematic changes in threshold level from trial to trial. Estimate variances using deviations from the means. Interpret your results.

2. Make a similar test of the observations made on the second day, estimating variances from the original measurements. If F proves to be significant, make t tests to determine where the genuine changes occur.

[1] Lindquist, E. F. Statistical analysis in educational research. New York: Houghton-Mifflin, 1940. Snedecor, G. W. Statistical methods. Ames, Iowa: Collegiate, 1937. Johnson, P. O. Statistical methods in research. New York: Prentice-Hall, 1949. McNemar, Q. Psychological statistics. New York: Wiley, 1949.

[2] See Baxter, B. Problems in the planning of psychological experiments. *Amer. J. Psychol.*, 1941, **54**, 270–280. Also, Fisher, R. A. The design of experiments. Edinburgh: Oliver & Boyd, 1935.

DATA 10*A*.—DATA IN A TWO-WAY CLASSIFICATION

Day	Trial			
	I	II	III	IV
1	24	19	21	24
	26	12	16	18
	21	17	17	22
	17	20	18	18
2	18	15	16	15
	19	15	19	19
	18	14	17	16
	17	12	14	18

3. Treat the entire table of data as a two-way classification problem. Make F tests to determine the significance of the three special sources of variance (between trials, between days, and interaction of trials and days). Interpret your results.

4. Take out each source of variance in Data 10*A* step by step, as was demonstrated in Table 10.7.

CHAPTER 11

TESTING HYPOTHESES

We have already emphasized the point that experiment and statistical method go hand in hand. The one supplements the other. The experiment directs our observations and yields data. By means of statistical methods, we can summarize those data, interpret them, and determine their reliability.

The best experiments are those that are set up to test the truth or falsity of some hypothesis. From previous experience, we believe a certain thing to be true, but it requires a crucial test to enable us to accept or to reject the hypothesis. If the result comes out one way, the hypothesis is probably correct; if it comes out another way, the hypothesis is probably wrong. The term "probably" is inserted because there is no such thing in science as absolute certainty. We are only more or less sure that the result points to one conclusion rather than to another.

The assurance of a conclusion may be of any degree of intensity from "doubtful" to "maybe" to "very likely" to "almost certain" to "practically certain." Statistical procedures give more definite meaning to those degrees of doubt and assurance. In this chapter, particularly, we shall be concerned with giving those concepts more exact meaning, so that we may be able to conclude whether certain outcomes of observations could perchance have arisen by accident or whether they point to something definitely not accidental.

Null Hypotheses

General Meaning and Application of a Null Hypothesis.—In the two preceding chapters we had incidental references to null hypotheses. Here we will see a number of other applications of them. We very properly say "null hypotheses" in the plural, for there are many ways of stating a null hypothesis, depending upon the nature of the experimental problem. In very general terms, this kind of hypothesis merely states that in an experimental situation, or even in a nonexperimental situation, whenever things are enumerated or measured it is assumed for the sake of argument that nothing but the laws of chance are operating. An illustration from experiments on extrasensory perception (ESP) is very suitable.

Suppose that an experiment with the Duke University ESP cards is properly set up to prevent the receiver from being influenced by any cues except possible telepathic stimulation. There are five different symbols on the cards, and in a thoroughly shuffled deck they should come up at random. As each one comes up and an experimenter reads it silently, the receiver makes his judgment. The card is returned to the deck, which is reshuffled, and the next one to be transmitted is selected. Starting with the hypothesis that there are no factors (*including* ESP) at work to determine the receiver's responses, we should expect in the long run an average success of 20 per cent right, or 1 in 5. If any receiver gives an excess of correct responses over and above 20 per cent, we still have to determine whether this excess is significant or whether it could have occurred by the processes of sampling in his limited number of trials. If the excess is one that could have happened as much as once in 10 times (one sample of this size out of 10 such samples), we should still say that the null hypothesis is quite plausible. We could not say that it is certainly established; but we would by no means give it up. Even if the excess over 20 per cent were one that could happen less than once in 20 samples, though we should be more skeptical of the null hypothesis, we should be unjustified in completely rejecting it. When so large a discrepancy as we obtained could occur by sampling less than once in 100 times, we customarily reject the hypothesis. We then say that it is highly implausible. In making this decision, there is only one chance in 100 that we have made an error.

But note that this does not automatically lead us to conclude that the alternative (ESP) hypothesis is true. It does tell us that something other than guesswork is going on, but it does not tell us what that "something other than guesswork" really is. If our experiment is designed so as to exclude all other possible factors than ESP in this case, then, having reduced the crucial experiment to an either-or proposition, *i.e.*, either laws of chance or ESP, and having overwhelming indication that the chance hypothesis is wrong, we can accept the ESP hypothesis as true. Unfortunately, the identification and control of all other factors favoring correct responses here is exceedingly difficult. But, in general, the establishment of an experimental fact depends upon it. We shall see shortly how a statistical test of the null hypothesis can be made for this type of experiment; but first let us consider some simpler cases.

Direct Determination of the Probable Validity of a Null Hypothesis.— Our first example is a simple psychophysical test situation. A student asserts that he can distinguish between two tones whose stimuli differ

only 2 cycles per second. That is his hypothesis: that he possesses genuine power to discriminate this difference in pitch. We doubt him, thus automatically adopting a null hypothesis. Out of 6 trials, how many pairs should we require him to judge correctly before we give up our hypothesis and yield to his? Our hypothesis implies that when he judges the pair of tones he might just as well flip a coin and report "second higher" for "heads" and "second lower" for "tails." We should expect him, by such guessing, to be correct half the time or 3 times out of 6. But how much of an excess over 3 correct judgments will it take to convince us that he is not merely guessing?

In a set of 6 trials, there are 7 possible outcomes—all the way from 6 down to 0 correct judgments. In Table 11.1 are listed all the 7 possibilities and the probability of each event's occurring by random sampling (chance). According to the probabilities involved in the situation, we should expect only *one* "score" of 6 in 64 samples; we should expect 6 "scores" of 5, 15 "scores" of 4, and so on. These expectations are according to the laws of probability.

TABLE 11.1.—EXPECTED OCCURRENCES AND PROBABILITIES OF SPECIFIED NUMBERS OF CORRECT JUDGMENTS IN MAKING SIX JUDGMENTS AT RANDOM

Number of correct judgments	Times expected in 64 sets of judgments	Probability of this number occurring in random sampling	Probability of as many or more occurring	Probability of as few or less occurring
6	1	1/64	1/64	64/64
5	6	6/64	7/64	63/64
4	15	15/64	22/64	57/64
3	20	20/64	42/64	42/64
2	15	15/64	57/64	22/64
1	6	6/64	63/64	7/64
0	1	1/64	64/64	1/64

The Use of Binomial Expansion.—A mathematical way of deriving the probabilities for the seven scores is to apply the expansion of the binomial $(\frac{1}{2} + \frac{1}{2})^6$. In tossing a coin there are two possible, independent, outcomes: head or tail. The theoretical probability of a head occurring is 1/2 and the probability of a tail is also 1/2. The general expression for the binomial is $(p + q)^n$, where n is the number of coins tossed. As in the case of proportions, $p + q = 1$, and 1 to any power also equals 1, so $(p + q)^n = 1.0$. The generalized binomial expansion is

$$(p+q)^n = p^n + \frac{n}{1} p^{(n-1)} q + \frac{n(n-1)}{1 \times 2} p^{(n-2)} q^2 + \frac{n(n-1)(n-2)}{1 \times 2 \times 3} p^{(n-3)} q^3 + \frac{n(n-1)(n-2)(n-3)}{1 \times 2 \times 3 \times 4} p^{(n-4)} q^4 + \cdots + q^n \quad (11.1)$$

in which p and q can have any positive values so long as $p + q = 1$.

Applied to the problem with 6 coins, ($n = 6$),

$$\left(\frac{1}{2}+\frac{1}{2}\right)^6 = \left(\frac{1}{2}\right)^6 + 6\left(\frac{1}{2}\right)^5\left(\frac{1}{2}\right) + \frac{6 \times 5}{1 \times 2}\left(\frac{1}{2}\right)^4\left(\frac{1}{2}\right)^2 + \frac{6 \times 5 \times 4}{1 \times 2 \times 3}\left(\frac{1}{2}\right)^3\left(\frac{1}{2}\right)^3 + \frac{6 \times 5 \times 4 \times 3}{1 \times 2 \times 3 \times 4}\left(\frac{1}{2}\right)^2\left(\frac{1}{2}\right)^4 + \frac{6 \times 5 \times 4 \times 3 \times 2}{1 \times 2 \times 3 \times 4 \times 5}\left(\frac{1}{2}\right)\left(\frac{1}{2}\right)^5 + \left(\frac{1}{2}\right)^6$$

$$= \frac{1}{64} + \frac{6}{64} + \frac{15}{64} + \frac{20}{64} + \frac{15}{64} + \frac{6}{64} + \frac{1}{64} = 1.$$

If the seven fractions are summed the result is equal to 1. The probabilities coincide with those in Table 11.1 for the various scores. The numerators give the expected frequencies of the scores 0 to 6 inclusive, when the total number of scores is 64.

Testing Deviations from Expected Values.—In determining whether the student's hypothesis about his acuity for pitch differences has much claim for acceptance, we are interested in how far his obtained score deviates from that to be expected by chance. A chance score in this situation would be 3 correct judgments out of 6. How much deviation from a score of 3 does he need to overthrow the null hypothesis? A score as high as 6 would be expected one sixty-fourth of the time. One chance in 64 would seem to be between the 5 per cent and 1 per cent levels so commonly applied. But remember that these standards are applied to deviations in *both* directions from the mean. The chance for a score of 6 is equal to that for a score of 0, which deviates an equal amount in the opposite direction. The probability of *either* event occurring is the sum of the probabilities for the two separately, or 2/64, or 1/32. Thus, if the student obtained a score of 6, we would still have some confidence in his claim, even though the probability is not much beyond the 5 per cent level.

An obtained score of zero would be interesting to interpret. From a common-sense point of view one might argue that such a score is positive evidence for lack of ability. But note that from a statistical standpoint a score of 0 is just as significant as a score of 6. We should be just as

inclined to reject the null hypothesis as when the score was 6. But the alternative inference would be different. Whereas a score of 6 would lead to some belief in real ability to judge differences in pitch, a score of zero would indicate something biasing the student in the direction of reversal of judgments. If this conclusion seems unreasonable, remember that with zero ability in this task a 6 is just as likely to occur as a 0, if judgments are not otherwise biased. An obtained score of 6 *can* be one of the events resulting from pure guessing. This should stress the importance of setting a high confidence level as a basis for rejecting a null hypothesis.

How about deviations of +2 or −2? Here we consider the odds of his having a score of 5 *or higher* or of 1 *or lower*. In Table 11.1 these probabilities are each 7/64, or combined, 14/64, or a little less than 1/4. Such an event would be less likely to occur than not, but the deviation would be much too small to be taken seriously. We would not reject the null hypothesis. A score as high as 4 or as low as 2 (which means all possibilities except a score of 3), each with a probability of 22/64 (see Table 11.1), gives a probability of 44/64, which means more often than not. On the whole, we would conclude from this line of reasoning that a test involving only six judgments would not be very decisive even when all judgments were correct. We should want more than six trials and we could determine the number of correct judgments required to justify rejection of the null hypothesis by a procedure like that described.

We consider next a case with a larger number of trials—a set of 10 true-false test items to which a student gives one of two alternative responses, one right and one wrong. How many more than 5 items must he do correctly for us to reject the hypothesis that he knows nothing about the subject matter of the examination and that he is merely guessing at random? The probabilities corresponding to the four highest scores are

TABLE 11.2.—EXPECTED OCCURRENCES AND PROBABILITIES OF SPECIFIED NUMBERS OF CORRECT RESPONSES TO 10 TRUE-FALSE TEST ITEMS

Number of correct responses	Expected number right in 1,024 sets	Probability of this number by chance	Probability of this number or higher	Probability of a like deviation in opposite direction	Probability of a like deviation in either direction
10	1	1/1,024	1/1,024	1/1,024	1/512
9	10	10/1,024	11/1,024	11/1,024	11/512
8	45	45/1,024	56/1,024	56/1,024	7/64
7	120	120/1,024	176/1,024	176/1,024	11/32

given in Table 11.2. These probabilities are derived in the same manner as those for the preceding problem, namely, from the application of the binomial equation. In this case the equation is $(\frac{1}{2} + \frac{1}{2})^{10}$.

From Table 11.2 we see that a score as extreme as 10 (a deviation of 5 from the expected or mean score of 5) could occur only once in 512 attempts. A score of 10 almost certainly indicates some knowledge or ability measured by the test, though we do not know just how much. A score as extreme as 9 could occur 11 times in 512 attempts or about 1 in 46 attempts. This would indicate probable knowledge or ability but not with very great assurance. A score of 8 could occur about once in 9 attempts and is consequently not at all fatal to the null hypothesis—the hypothesis of no knowledge or ability in the area sampled by this test.[1]

Departures from Random Conditions.—In applying such tests of the null hypothesis to any practical situation such as this, however, it must be kept in mind that we are assuming that in the event of complete ignorance the examinee will guess purely at random. Experience tends to show that in the absence of knowledge human beings do not always guess or respond at random. They exhibit patterns of responses or pattern habits. With biases such as this in the picture, hypotheses based upon chance distributions must be made with great caution and sometimes are precluded. The presence of bias cannot be easily detected, but one evidence of it would be a "significant" deviation in an unreasonable direction, as when in a guessing situation a statistically significant number of *wrong* judgments or responses occurs. Goodfellow has shown in connection with "experiments" on telepathy over the radio, for example, when an audience made five successive guesses of "black" versus "white" there are a number of common sequence patterns.[2] Alternations occur less frequently than one would expect by chance; runs are avoided; and certain initial responses may be favored, sometimes in response to an incidental cue that an experimenter might well overlook.

The presence of such nonrandom effects is bothersome, but there are experimental controls that may help to prevent them. There is probably enough randomness under a wide range of behavior to make possible a very profitable use of the statistical tests that depend upon it.

[1] For a discussion of the problems of testing whether "runs" of the same response are of sufficient length to justify rejection of the null hypothesis, see Grant, D. A., New statistical criteria for learning and problem solution in experiments involving repeated trials. *Psychol. Bull.*, 1946, **43**, 272–282.

[2] Goodfellow, L. D. The human element in probability. *J. gen. Psychol.*, 1940, **23**, 201–205.

Hypotheses Based upon the Normal Curve.—In the previous illustrations, we actually counted up the total number of possible outcomes and also the number of times certain outcomes would be expected, and from these we obtained directly the probabilities that the null hypothesis was incorrect. There are other instances, when the number of responses we deal with is quite limited, in which a similar counting of cases can be done and the probability of extreme deviations from chance can be derived. When the number of possible outcomes is not small, however, this counting of cases, or even algebraic computations of permutations and combinations, is much less efficient than other methods that will be described next.

In a certain elementary-psychology laboratory experiment, we have the problem to determine whether students can perceive from photographs

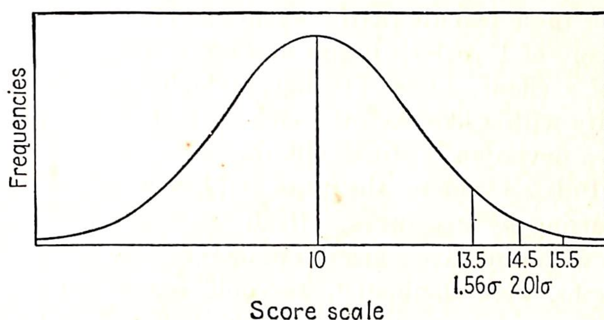


FIG. 11.1.—Standard-score distance from the hypothetical mean of the integral scores 14, 15, and 16 (correct judgments out of 20) when each judgment has an even chance of being right or wrong on the hypothesis of complete ignorance.

whether or not a man has been convicted of crime. Pictures of 20 pairs of men matched for certain qualities are exhibited, and the student judges which of the two is the criminal. The null hypothesis calls for 10 correct responses, provided that only random guessing accounted for the score. How large an excess is indicative of actual perception or of something other than chance?

To solve this problem, we do not resort to counting up the probabilities of as many as 20, 19, 18, etc., or more correct responses. Rather, we assume that each set of 20 judgments is a sample and that such samples would have a mean of 10, and a standard error of this mean will be the *SE* of a frequency, which equals $\sqrt{Npq}$ (see formula 9.19). We also assume a normal distribution of the samples of frequencies. For this problem, N is 20, p is .5, and q is .5. The σ_f is therefore $\sqrt{20 \times .5 \times .5} = 2.236$. The distribution of these frequencies is shown in Fig. 11.1, with a mean of 10 and a σ of 2.236. We are now ready to ask about the probability of a randomly determined score being as high as X or higher. For example, would a score of 14 be significantly in excess of the expected score of 10?

At first thought, this excess is 4 units above the mean of the distribution. But remember that a score of 14 is customarily one that occupies the interval from 13.5 to 14.5. A score of "14 or above" in this case therefore takes in all the normal curve above the point 13.5. It is a different matter to ask what is the area under the normal curve above the point 14.0 and to ask what is the area under the curve for a score of 14 or above. The deviation of the lower limit of this score from the mean is 3.5 units. Dividing this deviation by σ, which is 2.236, we have a t equal to 1.57.[1] Going to the probability table (Table B) with this standard score, we find the proportion of area above the point 13.5 to be .0583. Remembering that a score of 6 could occur as often as one of 14, and the probability of a score of 6 or below would also be .0583, the combined probability for these two alternative events is .1166, or about 12 chances in 100. A score of 15, which begins at 14.5, is 2.01σ above 10, and the probability of a chance score this high or higher is .0223. Combining this probability with a like one for a chance score of 5 or below, we have .0446. Such a deviation is just significant.

A score of 16 is 2.46σ above the mean and has only about 15 chances in 1,000 of occurring by guesswork. If all secondary cues, *i.e.*, cues not having to do with objective signs of criminality versus noncriminality in the photographs, were eliminated, we could conclude that the student who earns a score of 16 probably has the ability to make this kind of discrimination. If, however, we had obtained 1,000 scores and only 7 (approximately) were this large, we should, on the basis of this much larger experience, revert to the null hypothesis. But when we are restricted to a single sample of 20 judgments, the statistical tests justify us in rejecting the hypothesis when the score is as high as 16.

How Large a Deviation Is Significant?—To return to the ESP problem, in 50 trials, when the probability of chance success is .20 and so the expected frequency is 10, the standard error of the frequency is

$$\sqrt{50 \times .2 \times .8} = 2.83$$

We could now test the plausibility of the null hypothesis in the face of different numbers of correct responses in excess of 10. But it might be more to the point to ask how large a score it would take to be significant and how large a score to be very significant.[2]

[1] Remember that t is a standard measure, and in a normal distribution may be used just as z is used.

[2] See discussion on p. 200 regarding the sampling distribution of p (which also applies to the distribution of f).

To be significantly in excess of 10, a score of X or larger could happen by chance only 2.5 per cent of the time. What point on the score scale comes at such a position? From the table, the t corresponding to this point is 1.96. This value times σ is 1.96×2.83 units on the score scale. This excess added to 10 gives us 15.5. Remembering that a score of 16 really begins at 15.5, we conclude that *at least* a score of 16 or higher is required to be significant of *anything* over guesswork. To be *very* significant, the tail probability is .005, t is 2.576, and the excess is 7.3. This gives a point of 17.3 on the score scale. In terms of whole numbers, it requires a score of 18 or better to be very significant and to cause us to reject the null hypothesis. A score of 25 or better (above 24.5 on the scale) is 5.12σ above the mean, and there is less than one chance in a million that so large an excess could occur by guessing alone. Such scores demand an explanation, but the explanation is not inevitably to be in

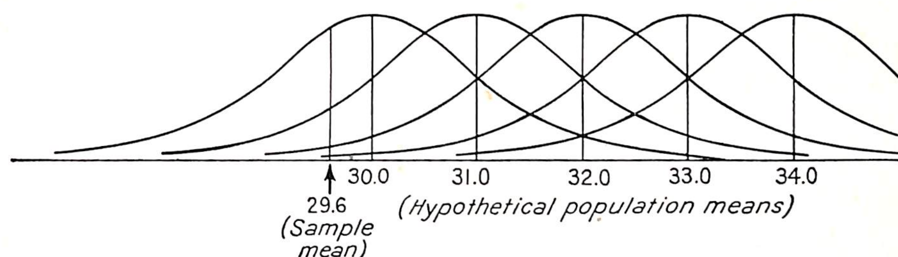


FIG. 11.2.—Hypothetical sampling distributions corresponding to various hypotheses concerning the population mean when the obtained or sample mean is 29.6.

terms of ESP unless other hypotheses have been adequately rejected by reason of rigorous and known control of experimental conditions.

Testing Different Hypotheses about the Population Mean.—It can now be shown, more appropriately than before, how, in the absence of other information, the obtained or sample mean is the most plausible value of the population mean. Let us return to the ink-blot test data used so many times in previous chapters. The mean of 50 scores was 29.6 in one sample, and the standard deviation was 10.45, from which the σ_M was estimated to be about 1.49. Let us choose several possible values that the population mean might have and in each case we will see how likely it is that a sample mean of 29.6 could then have arisen by random sampling. Figure 11.2 shows five hypotheses concerning the population mean: that it is, in turn, 34.0, 33.0, 32.0, 31.0, and 30.0.

Consider, first, the hypothesis that is farthest from the sample mean, namely, a hypothetical population mean of 34.0. This calls for an assumed sampling distribution with a mean of 34.0 and a standard deviation of 1.5. A sample mean of 29.6 deviates 4.4 from this hypothetical

mean. This deviation gives a t of $4.4/1.5 = 2.95$. What is the probability of a deviation as large as this occurring by random sampling? It is twice the area under the tail of the normal curve (since this sample qualifies as a "large" one). From Table B, we find the area in one tail beyond a z of 2.95 to be .0016. The area in both tails is .0032, from which we conclude that in only about 32 cases in 10,000 could a deviation as large as 4.4 units occur by random sampling. We therefore reject this hypothesis with considerable confidence.

The next hypothesis is for a population mean of 33.0, which gives a deviation of 3.4 for the sample mean and a t of 2.28. The area under the unit normal curve beyond this point is .0113. Twice this area is .0226. We can reject this hypothesis with only about 2 chances in 100 of being wrong. If we hypothesize a population mean of 32.0, the deviation is 2.4, t is 1.61, and the tail area is .0537. The chances for such a random deviation are more than 10 in 100. If we hypothesize a population mean of 31.0, the deviation is 1.4, t is .94, and the area beyond this t (in both directions) is .348. There are 348 chances in 1,000 that so large a deviation as 1.4 units could occur by random sampling. We do not reject the hypothesis that the population mean is 31.0. To go one step closer to the sample mean with our hypothesis, let us choose 30.0 as the population mean. This leads to an area of .394 in one "tail" and .788 in the two. The odds are 788 in 1,000 that a deviation as large or larger than that of 29.6 from 30.0 could have occurred by random sampling. Thus, as we approach the sample mean closer and closer with our hypothetical population mean, the odds keep increasing, which indicates that the plausibility of the hypothesis increases. The maximum plausibility would be reached when the hypothesis is 29.6, in other words, when it coincides with the sample mean.

In this discussion, we have omitted reference to the customary 5 per cent and 1 per cent levels. We could choose hypothetical population means such that the deviation of the sample mean from them would give t's at those particular levels. These deviations are known as *fiducial limits*. They mark off the limits of all hypotheses that give less than 5 per cent or 1 per cent degrees of confidence of rejection. All hypotheses of population means differing more than about 2.9 (which is 1.96 times σ_M) from the sample mean (in the ink-blot data) can be rejected with at least the 5 per cent degree of confidence. There are less than 5 chances in 100 of being wrong in so doing. All hypothetical means differing more than 3.8 can be rejected with confidence at the 1 per cent level. These interpretations of σ_M seem roundabout but in all logical accuracy this is how they are best made, though they are not so simple as those in Ch. 9.

How Large a Sample Is Necessary for Significant Deviations from Null Hypotheses?—We have already raised and answered the kind of question that asks for a given size of sample how large a discrepancy is necessary for significant and very significant deviation from a null hypothesis. Here we face a little different kind of question. We let our relative excess remain constant and ask how large N must be in order for that same size of discrepancy to reach the critical levels.

In a survey like the Gallup poll, for example, one would constantly be faced with the question of how large a sample to obtain; how many interviews to make; how many responses to a stimulus to record. That mere numbers in a sample as such are not sufficient to guarantee predictive ability was brought home to us decisively by the unhappy *Literary Digest* poll of 1936. Though the votes sampled ran into the millions, the voters who really determined the outcome of the presidential election were not adequately represented in the sample. A good poll sees to it that every kind of group of voters where group differences count at all are proportionately represented in the poll. When this is accomplished, it is surprising to the uninformed person how small a total sample can yield a valid predictive index. In other words, it is not so much enormous numbers that count as how the sample is made up.

Let us assume that our sample is properly made up with good representation.[1] Let us assume an issue where majority vote is decisive. Our null hypothesis is then 50 per cent or a proportion equal to .50. We ask first how large a sample is needed to give us confidence that an obtained vote of 55 per cent in favor of the proposition means a majority sentiment in that direction and did not occur by random sampling from a population that is on the fence. If a discrepancy of as much as 5 per cent is to be significant in our accepted meaning of the word, 5 per cent must deviate as much as 1.96σ from the mean of a normal distribution. In terms of proportions, the deviation is .05; how large must σ_p be? Obviously it must be such that .05 is 1.96 times σ. σ_p is therefore equal to .05/1.96, which equals .0255. The formula we need is

$$N = \frac{pq}{\sigma^2_p} \qquad \text{(Size of sample needed for significant deviation)} \qquad (11.2)$$

We know p and q and σ_p already. Substituting them in the equation, we have

[1] For the case of stratified sampling that is usually applied in public-opinion polling, modifications in line with standard-error formulas that fit that situation should be applied (see Ch. 9) rather than the general one for completely random sampling that is illustrated here.

$$N = \frac{.5 \times .5}{.0255^2} = \frac{.25}{.00065025} = 384$$

to the nearest whole number. It is therefore a 19 to 1 bet that when a vote comes out with 55 per cent in favor of an issue in a sample of 384 that the population sampled is not evenly divided on the question.

But where much is at stake, we should not be satisfied with these odds against the null hypothesis. We might ask how many votes need to be sampled to assure us of a *very* significant deviation. In this case, the excess of .05 must be at 2.576σ from the mean. The σ_p must be .05/2.576, which equals .0194. Applying formula (11.2) to determine N, we have

$$N = \frac{.5 \times .5}{.0194^2} = \frac{.25}{.00037636} = 664$$

Thus, in a sample of 664 interviewees, a majority vote of 55 per cent would be regarded as very significant. The odds would be 99 to 1 that the sentiment of the population sampled is not evenly divided on the issue. And since the deviation is in the direction favoring the issue, we strongly expect future outcomes to be in the same direction, but we do not know by how much.

The sizes of samples just found are surprisingly small in view of the enormous populations that vote on national issues and whose sentiment they may be expected to estimate. The reason is that we have allowed a rather wide margin of .05 as the deviation from null hypothesis. In dealing with more vital issues, where close elections are concerned, excesses of .01 or less may be decisive. If we are interested in the sizes of sample required to give significant and very significant indications when the vote is .51 to .49, the *SE* of the proportion must be one-fifth as large as it was for a .55 to .45 division. If σ_p is one-fifth as large, σ^2_p is one twenty-fifth as large. In this particular problem, the numbers to be substituted in formula (11.2) are now the same except that the denominator is one twenty-fifth of its former size. This makes N twenty-five times as large as before.

For a deviation of .01 to be significant now, N must be 9,600 and to be very significant it must be 16,600, these numbers being 25 times 384 and 664 respectively. Samples of this size would give us great assurance, granting random sampling, that the sentiment is in the direction indicated. On many issues, of course, the sentiment is more unevenly balanced than .55 and .45. And, again, when we are interested in significance of changes in sentiment, we have a revision of our problem, for then we are dealing with differences among proportions, a kind of problem to which we now turn.

Chi Square

Consider the data in Table 11.3, where we have a comparison of two samples; one is of 206 young American males who when they were in school had been regarded as feebleminded in terms of *IQ*. Their *IQ*'s were in the range 60–69. The other group is of 206 men of similar age (in the twenties) of *IQ*'s near 100.[1] These two groups had been compared with respect to a number of variables, one of which was marital status. At the time the study was made, the proportions married in the two groups were .539 and .408 for the normal and feebleminded groups, respectively. Is this difference significant?

Chi Square as a Test of a Null Hypothesis.—To answer the question just asked we might resort to the *t* test described for such a purpose in Ch. 9. We have an alternative procedure in the statistic known as chi square. It can be applied to this problem and also to a wider range of problems where a *t* test cannot be made. We will illustrate it first with this simple case. For comparison with the *t* test, it might be said that the obtained difference (.131) is 2.66 times its standard error, indicating a confident rejection of the hypothesis of no difference.

Besides assuming no real difference in proportion married, we can formulate a null hypothesis in another way to fit the chi-square approach. We make the same basic assumption in both cases: we adopt the idea that the two groups arose by random sampling from the same population. We then ask the question, "If this be true, how likely is it that the distribution of cases like those obtained could depart as much as they do from a random or chance distribution?" The four frequencies, in the four cells of Table 11.3, are 111, 84, 95, and 122. There seems to be some tendency for a concentration of cases in two cells: married-normal and unmarried-feebleminded. On the face of it, this looks like a meaningful departure from a random distribution.

If the distribution *were* random, what would it look like? We must determine the answer to this question, for that is the distribution called for by the null hypothesis. We do this entirely from the marginal totals, *i.e.*, the sums of rows and of columns. We take these values to be fixed, not having any better information about the genuine proportions of married versus unmarried and of feebleminded versus normal in the general population. Actually, we are not very much concerned about those proportions. We are abstracting two variables for study—marital status and intelligence—and we are interested in those qualities as such.

[1] From Baller, W. R. A study of the present status of adults who were mentally deficient. *Genet. Psychol. Monogr.*, 1936, **18**, 165–244.

The experiment naturally brings about certain restrictions in order to exert certain experimental controls. It is within the framework of the investigation that we are talking about a population. This fact is the logical justification for accepting the marginal sums as fixed. Within those limitations there is considerable freedom of variability.

The random distribution which will describe the null hypothesis is composed of four cell frequencies corresponding to those already mentioned: 111, 84, 95, and 122. We use the symbol f_o to stand for these observed frequencies and the symbol f_e to stand for the expected frequencies. How shall we find the expected frequencies?

Computation of Chi Square in a Contingency Table.—Reference to Table 11.3 shows that of the total sample of 412, 195 were married and

TABLE 11.3.—A COMPARISON OF MEN OF NORMAL *IQ* WITH FEEBLEMINDED MEN WITH RESPECT TO MARITAL STATUS

Marital status	Normal	Feebleminded	Both
Married	111	84	195
Unmarried	95	122	217
Total	206	206	412

TABLE 11.4.—THE EXPECTED NUMBERS OF MARRIED AND UNMARRIED MEN IN THE NORMAL AND FEEBLEMINDED GROUPS HAD THERE BEEN NO DIFFERENCE BETWEEN THE TWO

Marital status	Normal	Feebleminded	Both
Married	97.5	97.5	195
Unmarried	108.5	108.5	217
Total	206	206	412

217 were not. The proportions are .4733 and .5267, respectively. These proportions we may take to describe the (assumed) single population. By random sampling, each group, normal and feebleminded, should show the same proportions—.4733 married and .5267 unmarried. These proportions of 206 (for the normal group) give 97.5 and 108.5 married and unmarried persons, respectively. These are products (206 × .4733) and (206 × .5267), respectively. Since the feebleminded group also numbered 206, it would be expected to have the same frequencies. The entire set of expected frequencies is given in Table 11.4. If we add the columns and rows we find that the sums are the same as for the observed frequencies. It is always well to check one's work in this manner.

TABLE 11.5.—DISCREPANCIES BETWEEN OBTAINED AND EXPECTED FREQUENCIES IN TABLES 11.3 AND 11.4

Marital status	Normal	Feebleminded
Married............	13.5	−13.5
Unmarried..........	−13.5	13.5

TABLE 11.6.—THE CELL- SQUARE CONTINGENCIES FOR THE COMPUTATION OF CHI SQUARE RELATIVE TO THE STUDY OF MARITAL STATUS AND INTELLIGENCE

Marital status	Normal	Feebleminded	Both
Married..............	1.87	1.87	3.74
Unmarried...........	1.68	1.68	3.36
Both...............	3.55	3.55	7.10

Computing Expected Cell Frequencies.—In a contingency table of any number of rows and columns, the principles of computing the expected cell frequencies can be illustrated by the limited 3 × 3 table shown in Table 11.7. Let the f's with double subscripts stand for the obtained frequencies. The sums of the rows are symbolized by Σf_a, Σf_b, Σf_c, etc., and the sums of columns by Σf_1, Σf_2, Σf_3, etc. The expected frequency for any cell in row r and column k can be found by the formula

$$f_e = \frac{(\Sigma f_r)(\Sigma f_k)}{N} \quad \text{(Expected frequency for a cell in row } r \text{ and column } k\text{)} \qquad (11.3)$$

TABLE 11.7.—SCHEMA AND SYMBOLS FOR COMPUTATION OF EXPECTED CELL FREQUENCIES IN A CONTINGENCY TABLE

Rows	Columns			Sums of rows
	1	2	3	
A	f_{a1}	f_{a2}	f_{a3}	Σf_a
B	f_{b1}	f_{b2}	f_{b3}	Σf_b
C	f_{c1}	f_{c2}	f_{c3}	Σf_c
Sums of columns........	Σf_1	Σf_2	Σf_3	N

Let Σf_r stand for a sum of any rows, *e.g.*, Σf_a, Σf_b, . . . , etc.
Let Σf_k stand for a sum of any column, *e.g.*, Σf_1, Σf_2, ...

Thus, the expected frequency corresponding to f_{b3} would be derived from the product (Σf_b) (Σf_3) divided by N. Hence, the expected frequency for row 1 and column 2 of Table 11.4 would be equal to

$$\frac{(195)(206)}{412} = \frac{40170}{412} = 97.5$$

Computing Cell Discrepancies.—Having the expected frequencies f_e, we now ask whether the observed frequencies f_o deviate from them sufficiently to cause us to reject the hypothesis of no difference. For each of the four cells of the table, we determine the discrepancy $f_o - f_e$. These discrepancies are listed in Table 11.5. It will be seen that except for algebraic sign they are all numerically the same. This outcome will be true of all fourfold tables of frequencies of this sort, whether the two groups compared have the same total numbers of cases or not. This fact can be used to give us short cuts in computation, as we shall see later.

The Cell Square Contingencies.—In the solution of chi square, we square each discrepancy, divide by the corresponding f_e, and sum all the ratios. The sum is chi square. In terms of a formula,

$$\chi^2 = \sum \left[\frac{(f_o - f_e)^2}{f_e} \right] \qquad \text{(General formula for chi square)} \qquad (11.4)$$

where the symbols have been explained above. Each cell provides a ratio of $(f_o - f_e)^2$ to f_e, which ratio has been called the *cell square contingency*. This is merely a convenient name, at present, but later (Ch. 14) it will be related to prediction procedures. For now, it can be said that chi square is the sum of the cell square contingencies in a contingency table (see Table 11.6).

The square of the discrepancy 13.5 is 182.25. In two cells, this is to be divided by 97.5, which yields 1.87. In the other two cells it is to be divided by 108.5, which yields 1.68. Summing twice 1.87 and twice 1.68, we have 7.10 as the value of χ^2.

Interpretation of a Chi Square.—The number 7.10 stands for the total amount of discrepancy between hypothesis and observation. Chi square can be small enough to allow us to accept the null hypothesis or to retain it with some doubt, or it can be large enough to lead us to reject the hypothesis with moderate or with positive assurance. Like Student's t ratio, it can be interpreted as being significantly or very significantly large, *i.e.*, of being so large that sampling alone could account for the results only once in 20 times, or once in 100 times, as the case may be.

Degrees of Freedom.—Tables of chi square (see Table E, Appendix B) enable us to decide the matter. But we must know the number of degrees

of freedom *df* before we can use the table. In a fourfold table such as we have here, there is only 1 degree of freedom.

Let us see how it is that we have only 1 degree of freedom. Remember that we have taken the row and column sums to be fixed. This injects considerable rigidity into a contingency table. The general rule applying to all contingency tables regardless of numbers of rows and columns is that the degrees of freedom equal the product of the number-of-rows-

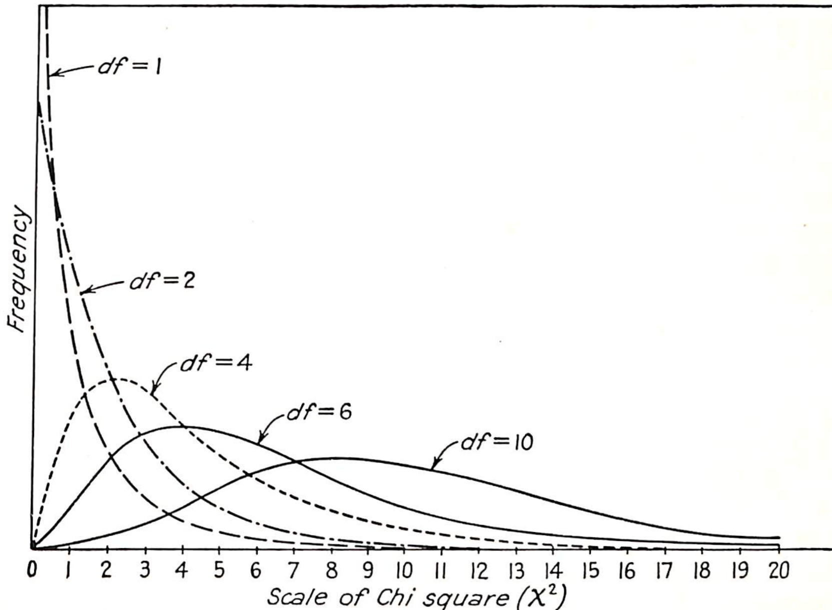


FIG. 11.3.—Sampling distribution of chi square for various degrees of freedom. (*After Lewis, D. Quantitative methods in psychology. Iowa City: The author*, 1948.)

minus-one and the number-of-columns-minus-one. If there are r rows and k columns, both r and k being greater than 1,

$$df = (r - 1)(k - 1) \quad \text{(Number of degrees of freedom in a contingency table of } r \text{ rows and } k \text{ columns)} \qquad (11.5)$$

In a 2×2 table, applying the formula, we would expect 1 degree of freedom. This is made reasonable by the following logic. Once we have chosen a single cell frequency, with the row and column sums being what they are, all the other cell frequencies are determined; are not free to vary. This is reflected, also, by the fact that there is only one value for the cell discrepancies.

The Sampling Distribution of Chi Square.—The importance of degrees of freedom can be seen in connection with Fig. 11.3, which shows the sampling distributions of chi square for a number of different degrees of

freedom ranging from 1 to 10. It is because of these known distributions that the tables for interpreting a chi square could be constructed. In general, distributions of this statistic are positively skewed, and the smaller the degree of freedom, the greater the skewness. As the number of degrees become large, this distribution approaches the normal curve in form. The distribution with 10 degrees of freedom is apparently not far from normal.

Use of the Chi-square Tables.—Is our chi square of 7.10 significant? Table E shows that when $df = 1$ the largest chi square given is 6.635. Right above this is the probability of .01, which means that a chi square as large as 6.635 or larger could occur by chance alone only once in 100 times. Our chi square of 7.10 is larger than 6.635 and therefore could occur in the same manner less than once in 100 times. We therefore regard it as very significant and reject the hypothesis of no difference between the two groups.

Relation of Chi Square to t.—When there is 1 degree of freedom in a contingency table, chi square is equal to t^2, or t is equal to chi, the square root of chi square. The square root of our chi square obtained for the marital data, namely 7.10, is equal to 2.66. This checks exactly with the t that was reported in an earlier paragraph. A t test and a chi-square test of the same statistics will therefore lead to the same inferences when there is 1 degree of freedom.

Chi Square When Frequencies Are Small.—When applying the chi-square test to a problem with 1 degree of freedom, particularly, it is important to exercise caution when samples are small. When any of the theoretical cell frequencies are small (some say as small as 50; in the opinion of the author 25 might be a more realistic limit), it is recommended that a correction known as *Yates's correction for continuity* (correction for *discontinuity* would be a better name for it) be applied. This correction consists of reducing by .5 each of the obtained frequencies that are greater than expected and of increasing by .5 each of the frequencies that are less than expected. This has the effect of reducing the amount of each of the discrepancies between observed and expected frequencies by .5 (without changing the expected frequencies). It is especially important to apply such a correction when the chi square turns out (without correction) just above the customary confidence limits. The effect of the correction is to reduce the size of chi square. The correction is needed here for the same reason it was suggested in connection with a distribution of scores (see p. 267). There it was pointed out that with a limited number of score units in the range, testing whether a given score was significant involved reducing a deviation by .5 unit in order to get to the point at its lower

limit. Frequencies, like scores, increase by whole units, whereas distributions of sampling statistics like t and chi square are on a continuous scale. When scores or frequencies are numerically large, a change of .5 unit is relatively unimportant, but when the number of either is limited, a change of .5 unit amounts to quite a percentage change.

There are lower limits to utilizable frequencies, even when Yates's correction is applied. Some authors say that a chi square should not be computed if any theoretical frequency is less than 10. Others, more generous, would compute chi square even when a theoretical cell frequency is as low as 2. A realistic limit is 5. If there are more than four cells in the contingency table, there is always the possibility of combining cells so as to avoid very low frequencies. An illustration or two of this will be shown later (see Table 11.10). There is probably nothing to be gained by applying Yates's correction when there is more than 1 degree of freedom, and under these conditions the correction becomes complicated.[1]

An Example of Yates's Correction.—In a public-opinion poll conducted not so many years ago, sentiment was sampled concerning attitudes toward radio newscasts.[2] Some 43 interviewees in one sample were asked the question, "Do you find it easier to listen to news than to read it?" The sample had been stratified into a higher and a lower socioeconomic status, 19 being in the former and 24 in the latter. The numbers responding "Yes" to the question in the two groups were 10 and 20, respectively. The problem to be investigated is whether there is a real difference between the two groups in their opinions on the question.

The data have been arranged in the usual manner in Table 11.8. The frequencies are all small enough, certainly, to call for Yates's correction. One obtained frequency is less than 5, but we pay more attention to the theoretical frequencies before we may decide to give up the idea of a chi-square test. The expected frequencies are given in Table 11.8. All of them exceed 5, so we may proceed with the test.

Let us carry through the test first *without* the correction and then with it to see what difference it may make in our conclusions. Without the correction, the cell deviations would all equal 3.26. This squared is 10.63. Applying formula (11.4) and solving, we find that chi square equals 4.76, which is significant between the 5 per cent and 1 per cent levels. With the correction applied, the cell deviation in all cells is 2.76 rather than 3.26, whose square is 6.72. With this solution, chi square

[1] Kelley, T. L. Fundamentals of statistics. Cambridge: Harvard University Press, 1947. P. 330.

[2] From Cantril, H. The role of the radio commentator. *Publ. Opin. Quart.*, 1939, **3**, 654–662.

becomes 3.42, which fails to reach the 5 per cent of significance. One would have much more confidence in the interpretation of the second outcome than the first. Not always will the correction make as much

TABLE 11.8.—COMPUTATION OF CHI SQUARE FOR RESPONSES OF TWO SOCIOECONOMIC GROUPS TO PREFERENCE FOR RADIO NEWS TO READING A NEWSPAPER

Response	Obtained frequencies			Expected frequencies		
	Socioeconomic group			Socioeconomic group		
	Higher	Lower	Both	Higher	Lower	Both
Yes	10	20	30	13.26	16.74	30
No	9	4	13	5.74	7.26	13
Both	19	24	43	19	24	43

difference and not always will it lead to a drastic change in the conclusion. When in doubt, use it.

Other Ways of Computing Chi Square in a 2 × 2 Table.—In a four-fold-table problem, since the discrepancy is the same for all cells, the formula for chi square can be written

$$\chi^2 = (f_o - f_e)^2 \sum \left(\frac{1}{f_e}\right) \quad \text{(Chi square in a } 2 \times 2 \text{ contingency table)} \qquad (11.6)$$

That is, chi square equals the common discrepancy squared times the sum of the reciprocals of the four f_e's. As applied to the marital-status problem

$$\begin{aligned} \chi^2 &= 13.5^2 \left(\frac{1}{97.5} + \frac{1}{97.5} + \frac{1}{108.5} + \frac{1}{108.5}\right) \\ &= 182.25(.01026 + .01026 + .00922 + .00922) \\ &= 182.25 \times .03896 \\ &= 7.10 \end{aligned}$$

If the data are arranged in a 2 × 2 table as shown in Table 11.9, another convenient formula for the computation of chi square is

$$\chi^2 = \frac{N(ad - bc)^2}{(a + b)(a + c)(b + d)(c + d)} \quad \text{(Alternative formula for chi square in a four-cell, } 2 \times 2 \text{ table)} \qquad (11.7)$$

Applied to the opinion-poll data,

$$\chi^2 = \frac{43[(10)(4) - (20)(9)]^2}{(30)(19)(24)(13)}$$
$$= \frac{43(40 - 180)^2}{(570)(312)}$$
$$= \frac{842,800}{177,840}$$
$$= 4.74$$

The answer is close enough to that obtained by the other method to be regarded as checking. Note, however, that this is the result without correction for continuity.

TABLE 11.9.—SYMBOLIC ARRANGEMENT OF DATA IN A 2 × 2 CONTINGENCY TABLE ILLUSTRATED BY THE PUBLIC-OPINION DATA

Variable I	Variable II: Higher	Lower	Both
Higher......	a	b	$a + b$
Lower......	c	d	$c + d$
Both.....	$a + c$	$b + d$	N

	Socioeconomic group: Higher	Lower	Both
Yes.........	10	20	30
No.........	9	4	13
Both.....	19	24	43

Chi Square in Other than 2 × 2 Tables.—The use of chi square is by no means limited to fourfold contingency tables. It can be applied with as few as two cells and with a much larger number. First, an example with only two frequencies to be tested.

In a Two-cell Table.—For this purpose let us use the polling data on preference for the radio. Combining the two socioeconomic groups together as representing a more heterogeneous population, we may be interested in knowing whether the population they represent is actually in favor of radio newscasts. The sample is so small that there may be some doubt. The frequencies are 30 in favor and 13 not. Could these frequencies have arisen from a population in which the division of opinion is really evenly divided? The null hypothesis for this purpose is a 50-50 division. This is an arbitrarily chosen hypothesis; we could have chosen some other, such as a 60-40 division of opinion.

With the 50-50 hypothesis chosen, the expected frequencies are 21.5 and 21.5, these being one-half of 43. The cell deviations or discrepancies ($f_o - f_e$) are 8.5, one positive and the other negative (without correction for continuity). The squared discrepancy is 72.25. Dividing this by f_e, which is the same in both cells, we get a squared contingency of 3.360 for each cell. For the two combined we get 6.720, or a chi square

of 6.72. This is significant just beyond the 1 per cent level. With so small a sample, we should apply the correction for continuity, in which case the cell discrepancies are 8.0 and squared they are 64.0. The cell-square contingencies are 2.977 and chi square is 5.95. This figure leads us to a modification of our confidence in a real difference, though it does not reverse our decision about rejection of the 50-50 hypothesis.

With a two-cell table, when expected frequencies are equal, as in the last illustration, the formula for chi square reduces to the simple form

$$\chi^2 = \frac{2(f_o - f_e)^2}{f_e} \quad \text{(Chi square in a two-cell table when expected frequencies are equal)} \quad (11.8)$$

Since, with one degree of freedom $t = \chi$, another formula for t, derived from (11.8) and applying in the same special but not uncommon situation, is[1]

$$t = \frac{f_1 - f_2}{\sqrt{f_1 + f_2}} \quad \text{(t test of departure of two frequencies from equality)} \quad (11.9)$$

where f_1 = the larger of two frequencies and $f_1 + f_2 = N$.

Applied to the polling problem,

$$t = \frac{30 - 13}{\sqrt{30 + 13}} = \frac{17}{\sqrt{43}} = 2.59$$

The square of this value is 6.71, which checks with the chi square obtained above, without correction for continuity. Correction for continuity would involve the use of the expression $(f_1 - f_2 - 1)$ in the numerator of formula (11.9) in place of $(f_1 - f_2)$.

Chi Square in Larger Tables of Frequencies.—To illustrate the application of chi square to a larger table, this time with a table of six cells, let us consider some more survey-of-opinion data.[2] This time the question was whether the radio listener agreed with the opinions expressed by a certain radio commentator, and the responses were tabulated as

[1] By a little algebra, it will be found that $(f_1 - f_2) = 2(f_o - f_e)$ and that

$$f_e = (f_1 + f_2)/2.$$

Equation (11.8) then becomes

$$\chi^2 = \frac{(f_1 - f_2)^2}{f_1 + f_2}$$

[2] Cantril, *op. cit.*

"Agree," "Disagree," or "Doubtful." The survey was made in two cities and we have the numbers responding in each way in both of them. The results are listed in Table 11.10.

TABLE 11.10.—A CHI-SQUARE SOLUTION IN A TWO-BY-THREE TABLE OF DATA ON OPINIONS EXPRESSING AGREEMENT OR DISAGREEMENT WITH A CERTAIN RADIO COMMENTATOR

Categories of response	Opinions in Syracuse	Opinions in Columbus	Both
Agree	73	22	95
Disagree	9	4	13
Doubtful	41	27	68
Totals	123	53	176

f_e Expected frequencies		$f_o - f_e$ Discrepancies		$(f_o - f_e)^2$ Discrepancies squared		$\frac{(f_o - f_e)^2}{f_e}$ Ratios	
Syracuse	Columbus	Syracuse	Columbus	Syracuse	Columbus	Syracuse	Columbus
66.4	28.6	+6.6	−6.6	43.56	43.56	0.66	1.52
9.1	3.9	−0.1	+0.1	0.01	0.01	0.00	0.00
47.5	20.5	−6.5	+6.5	42.25	42.25	0.89	2.06
123.0	53.0	0.0	0.0	—	—	1.55	3.58

The derivation of the expected frequencies was carried out with the application of formula (11.3). From here on, the work as recorded in Table 11.10 is just as we have done previously. The sum of the square contingencies is 5.13. The degrees of freedom (according to formula 11.5) are $2 \times 1 = 2$. For 2 degrees of freedom the tables of chi square show that it requires a chi square of 5.991 to be significant at the 5 per cent level. Our chi square falls below this level, so there is no really convincing reason to doubt that the two populations sampled are alike on the question at issue, though there are less than 10 chances in 100 that a chi square as large or larger could have arisen by chance.

The small expected frequencies in Table 11.10 should raise some question concerning the need for corrections.

Combining Columns or Rows.—As a matter of fact, we have one expected frequency lower than 5. If we decide that it is too risky to solve the problem with so small an f_e, there is one thing we could do. Incidentally, it happens in this particular problem that the squared discrepancy

$(f_o - f_e)^2$ was practically zero for the cell in which f_e was smallest, so that this cell makes no contribution to chi square. It is a situation in which a very small f_e is combined with a relatively large squared discrepancy that is serious, for then the cell's contribution to chi square is unduly large and yet of doubtful stability or meaningfulness.

If we had combined the "Disagrees" with the "Doubtfuls" in this problem, we should have had observed frequencies of 50 for Syracuse and 31 for Columbus, with expected frequencies of 56.6 and 24.4, respectively. We can combine both observed and expected frequencies after the latter have been computed in uncombined form. After this kind of a combination is made, the size of chi square is likely to be smaller than before, though not always. Even though it is smaller, the number of degrees of freedom is also reduced and the significance limits are accordingly smaller, so that the chances of a significant departure of data from the null distribution are presumably about the same as they were.

Chi Square in Testing the Hypothesis of Normal Distribution.—One convenient use of chi square is in testing whether or not a set of observed frequencies in a frequency distribution could probably have arisen from a normally distributed population. The procedure is carried out in much the same manner as with frequencies in this chapter. Expected frequencies are estimated as was illustrated in Ch. 7, particularly in Table 7.1. The discrepancies between observed and expected frequencies are squared, divided by the f_e's, and these ratios are summed to give χ^2. The number of degrees of freedom is the number of class intervals less three. One degree of freedom is lost in computing the mean, one in computing the standard deviation, and one for N, the size of sample. All three of these statistics are used in deriving the expected frequencies, since the three of them describe the particular normal curve that comes closest to the data.

At the tails, where f_e's are small (less than 5), two or more class intervals should be grouped together as was suggested in recent paragraphs for other data. The interpretation of the result is made according to precedents already repeated, and the hypothesis of normality is accepted or rejected according as χ^2 is small or large.

Using the data of Table 7.1, with the expected and obtained frequencies already given, we will make the chi-square test. First, to get rid of very small tail frequencies we will combine the four classes at the upper end of the distribution and the three at the lower end. The results of this are shown in Table 11.11. The next steps are carried out, as shown, with a resulting chi square equal to 3.68. With 5 degrees of freedom, a chi square of 11.07 is required for significance at the 5 per cent level. From the chi-square table, we find by interpolation that about 60 per cent of the chi

TABLE 11.11.—A CHI-SQUARE TEST OF THE NORMAL-DISTRIBUTION HYPOTHESIS APPLIED TO A FREQUENCY DISTRIBUTION OF SCORES

(1) Scores	(2) Original grouping		(3) Regrouped frequencies		(4) Cell discrepancies $f_o - f_e$	(5) Cell discrepancies $(f_o - f_e)^2$	(6) Cell square contingencies $\frac{(f_o - f_e)^2}{f_e}$
	f_o	f_e	f_o	f_e			
44–46	0	0.2					
41–43	1	0.8					
			10	8.2	+1.8	3.24	0.395
38–40	4	2.2					
35–37	5	5.0					
32–34	8	9.0	8	9.0	−1.0	1.00	0.111
29–31	14	13.3	14	13.3	+0.7	0.49	0.037
26–28	17	15.8	17	15.8	+1.2	1.44	0.091
23–25	9	15.1	9	15.1	−6.1	37.21	2.464
20–22	13	11.7	13	11.7	+1.3	1.69	0.144
17–19	8	7.2	8	7.2	+0.8	0.64	0.089
14–16	3	3.6					
11–13	4	1.5	7	5.6	+1.4	1.96	0.350
8–10	0	0.5					
Σ	86	85.9	86	85.9	+0.1		$\chi^2 = 3.681$

squares from data such as these could be as large as 3.68 or larger. We conclude that the obtained frequency distribution fits the normal form quite acceptably. We can well tolerate the idea that the population from which it came is normally distributed on the measurement scale. In making this chi-square test, it is important that Σf_e shall approach N very closely.

Exercises

1. Suppose that we ask an observer to arrange a series of weights in rank order from lightest to heaviest, the differences being very small. If he places them in perfect rank order, what is the probability that he could have done so by sheer guessing? No matter how many weights ranked, there is only one correct way of doing this. The total number of ways the observer could have arranged each number of weights is given below:

Number of weights...........	3	4	5	6	7
Number of orders............	6	24	120	720	5,040

Which perfect orders would be regarded as "not significant," "significant," and "very significant"? State the probabilities of perfect orders by chance.

2. An observer knows that he will hear one of three similar speech sounds. He is given the three in chance order in a total of 30 trials. How many correct judgments must he give before we regard his success as significant and as very significant?

3. Suppose that the observer in Exercise 2 were given 48 trials. How large a score is significant, and how large a score is very significant?

4. A certain examination includes 40 items, each item with four alternative responses. How large a score must a student earn before you feel that he probably knows something about the content of the examination? Before you feel that he undoubtedly knows something about it? Would you feel absolutely sure that he knows something about the content if he made a score of 35? Discuss.

5. In a test of five-response items, how many items would you need to include in order to feel sure that a score of 30 per cent right indicates knowledge of the content? How large must the test be if a score of 25 per cent right is to indicate knowledge beyond a reasonable doubt? Tell how you have interpreted "sure" and "beyond reasonable doubt."

DATA 11*A*.—NUMBER OF PERSONS IN TWO GROUPS, DEPRESSED AND NOT DEPRESSED IN TEMPERAMENT, WHO RESPONDED IN EACH OF THREE CATEGORIES TO THE QUESTION, "WOULD YOU RATE YOURSELF AS AN IMPULSIVE INDIVIDUAL?"

Group	Yes	?	No	Totals
Depressed	72	45	133	250
Not depressed	106	35	109	250
Totals	178	80	242	500

6. Is there a significant difference in Data 11*A* between the numbers of "Yes" responses? Present statistical proof.

7. Is there a significant difference between the two groups in the number of "?" responses? Explain.

8. Is there a significant difference in the two groups with regard to all three response categories taken together? Determine this by computing chi square.

9. State a number of null hypotheses that might be applied to Data 11*B*.

DATA 11*B*.—NUMBERS OF TWO GROUPS DIFFERING IN ABILITY WHO PASSED A CERTAIN TEST ITEM

Group	High group	Low group	Both
Passed	62	48	110
Failed	38	52	90
Both	100	100	200

10. Do both groups together in Data 11*B* show a significant deviation from a chance situation of passing and failing? Explain.

11. Is there a significant difference between the high and low group in terms of the numbers passing the item? Explain. Can you predict from this result whether there would be a significant difference between numbers of failures in the two groups? Explain.

12. Find a chi square for Data 11*B* in as many ways as you know how. Interpret your results.

13. In Data 11*A*, combine the "Yes" and "?" responses, and compute chi square for the fourfold table. Compare your results with those in Exercise 8.

CHAPTER 12

TEST SCALES AND NORMS

In this chapter we take a short departure from fundamental statistics but we apply statistical ideas to the problems of measurement by means of tests. Heretofore, we have accepted raw test scores as if they were measures of psychological or educational variables on scales of equal units. Strictly speaking, this assumption is necessary before we are justified in applying many of the statistical procedures, including arithmetic mean, standard deviation, and correlation coefficient. When a test is composed of many items, when it covers a relatively wide range of values, and when it is of an appropriate level of difficulty for the population examined, this assumption is fairly sound. Now, however, we must examine the question of measuring scales, not so much for their suitability in meeting the assumption of equal units but for the very practical reason of comparability and meaningfulness of scales.

Why Common Scales Are Necessary.—The chief reasons for dissatisfaction with most raw-score scales are their lack of meaning and their lack of comparability. Aside from a few tests that yield scores in terms of physical stimulus values (such as tests of sensory acuity) or of response values (such as time, distance, or energy values), most tests yield numerical values that have no particular significance. There was a time (unfortunately it still is not entirely in the past) when scores were given in terms of percentages. The tradition of grading examinations in terms of percentage of right answers still has popular appeal, in spite of the many experimental demonstrations that such percentages are neither accurate nor meaningful. The method gave a feeling (definitely falacious) of having some kind of an "absolute" measure of the individual. It is difficult for even the better informed student to free himself from this traditional thinking, even when he has given up the operations it implies.

If modern psychology and education have taught anything about measurement, they have amply demonstrated the fact that there are few, if any, absolute measures of human behavior. The emphasis has shifted from the search for absolute measures to an emphasis upon the concept of individual differences. The mean of the population has become the reference point; and out of the differences between individuals has come the basis for scale units. Even when the test happens to yield such objective scores

as those in time, or space, or energy units, it is sometimes doubted that such units, though undoubtedly equal from a physical point of view, really represent equal psychological increments along scales of ability or talent. These considerations, among others, send us in search of more rational and meaningful scales of measurement for behavior events.

In addition to the more theoretical demands just mentioned, there is the very practical consideration that scales for different tests should be comparable. The most obvious need for comparable scales is seen in educational and vocational guidance, particularly when profiles of scores are utilized. A profile is intended to give a picture of an individual. We would hardly bother to prepare one for an individual if we did not expect to make very direct comparisons of the person's levels in different traits. The comparisons of trait positions for the same individual would be misleading, if not worthless, if there were not at least reasonable comparability of levels for different scores going under the same numerical value. No informed person would think of using raw scores as a basis of making direct comparisons among an individual's positions with respect to trait variables. Conversion of raw scores to values on some other, common, scale is essential. The use of centile rank positions was mentioned in an earlier chapter (Ch. 6). Centile values are suitable to the extent that they do make possible comparable values for different tests, they do use the mean (or median) as the main reference point, and they are easily understood by the layman. They serve their best purpose when measurements must be interpreted to the layman. But, for reasons which were stated earlier (Ch. 6), centile values have limitations which make them fall short of full usefulness to those who expect something more of measurements. Centiles, after all, are rank positions and do not represent equal units of individual differences. It is possible to have scales that probably provide units of equal size as well as comparability of means, dispersions, and form of distribution.

Some Common Derived Scales.—The chief interest in what follows will be in such scales—those which achieve comparability of means, dispersions, and form of distribution. We will not go into the very popular mental-age concept or the *IQ* scale. As simple as those ideas may be, the achievement of a battery of tests which will meet the requirements of age equivalents and appropriate distributions of *IQ* involves statistical problems of an intricate nature which we cannot go into. Treatment of these problems may be found in references to McNemar and to Marks.[1] The three kinds

[1] McNemar, Q. The revision of the Stanford-Binet scale. Boston: Houghton Mifflin, 1942; Marks, E. S. Sampling in the revision of the Stanford-Binet scale. *Psychol. Bull.*, 1947, **44**, 413–434.

of scales to be discussed here are the standard-score scale, the *T* scale, and the C scale. Their application to derivation of test norms and profile charts will be given attention. The treatment will be kept at a rather elementary level, emphasizing basic concepts. For a more advanced treatment of some of these problems the reader is referred to a monograph prepared by Flanagan.[1]

STANDARD SCORES

An Example of the Need for Comparable Scores.—A concrete example will illustrate some of the ideas expressed above. A student earns scores of 195 in an English examination, 20 in a reading test, 39 in an information test, 139 in a general scholastic-aptitude test, and 41 in a nonverbal psychological test. Is he therefore best in English and poorest in reading? Could he perhaps be equally good in all the tests? From the raw scores alone, we can answer neither of these questions nor many others that could be legitimately asked. This student's (student I) five scores just cited will be seen listed in column (4) of Table 12.1. Knowing the means of

TABLE 12.1.—A COMPARISON OF STANDARD SCORES WITH RAW SCORES EARNED BY TWO STUDENTS IN FIVE EXAMINATIONS

(1)	(2)	(3)	(4)		(5)		(6)		(7)	
Examination	Mean	Standard deviation	X Raw scores		x Deviations		z Standard scores		$z - M_z$ Deviations in standard scores*	
			I	II	I	II	I	II	I	II
English	155.7	26.4	195	162	+39.3	+ 6.3	+1.49	+0.24	.98	.66
Reading	33.7	8.2	20	54	−13.7	+20.3	−1.67	+2.48	2.18	1.58
Information	54.5	9.3	39	72	−15.5	+17.5	−1.67	+1.88	2.18	.98
Scholastic aptitude	87.1	25.8	139	84	+51.9	− 3.1	+2.01	−0.12	1.50	1.02
Psychological	24.8	6.8	41	25	+16.2	+ 0.2	+2.38	+0.03	1.87	.87
Sums	....	....	434	397	......	......	+2.54	+4.51	8.71	5.11
Means	....	....	...	...	......	......	+0.51	+0.90	+1.74	1.02

* Disregarding algebraic signs.

students in the five tests helps some, since they serve as norms or comparable zero points. The means are listed in column (2). We now see that the student is well above average in English and in scholastic aptitude and is somewhat below average in reading and information, just as the

[1] Flanagan, J. C. Scaled scores. New York: Cooperative Test Service, 1939.

numbers seem to indicate at their face value. The second student, whose raw scores are also in column (4), is numerically highest in the same two and lowest in the same three. When we consider the averages again, however, we find that student II is only about average in English, in scholastic aptitude, and in the psychological test, but he is above average in reading and in the information test.

When a student is above the mean in two tests, in which one is he actually superior? Student I is 39.3 points above the mean in English and 16.2 points above the mean in the psychological test [see column (5) of Table 12.1]. Is his superiority in English really greater than his superiority in the psychological test? Student II is 20.3 points above the mean in reading and 17.5 points above the mean in information. Is he about equally superior in the two tests?

And how do the two students compare? The superiority of student I is apparent in three tests (English, scholastic aptitude, and psychological) and that of student II, in the other two tests. This we can tell from the raw scores. But suppose the two were competing for a scholarship at a university; which one, if there is to be a choice between the two, should win? The totals of the five scores are 434 and 397, in favor of student I. Granting that the five different abilities are equally important, have we done justice by comparing sums of raw scores? Should we be justified in finding an average of each student's five raw scores?

Suppose that we were interested in determining which student is the more consistent in his abilities, as shown by these five tests, and which one has the greater variability within himself. Would a comparison of the average deviations or standard deviations of the five raw scores give us the answer? As the reader has probably guessed, the reply to most of these questions is in the negative. We are extremely limited in making direct comparisons in terms of raw scores for the reason that raw-score scales are arbitrary and unique. We need a common scale before such comparisons as we have called for can be made. Standard scores furnish one such common scale.

The Nature of a Standard-score Scale.—A standard-score scale is one that has a mean of zero and a standard deviation of 1.0. The unit of the scale might be taken as 1σ, or as 0.1σ, or any other arbitrary fraction of the standard deviation. An illustration of the conversion of a raw-score scale into a standard scale is shown in Fig. 12.1, *A*, *B*, and *C*. Distribution *A* is based upon the original or raw scores. The mean is 80 and standard deviation is 14.0. The distribution is obviously somewhat negatively skewed.

As we have previously seen, a standard score z is derived from a raw score X by means of the formula

$$z = \frac{X - M}{\sigma} = \frac{x}{\sigma} \quad \text{(Standard score } z \text{ corresponding to a raw score } X \text{ and to a deviation } x\text{)} \qquad (12.1)$$

An intermediate step between the raw-score scale and the standard-score scale is the deviation $X - M$, or x. This step is illustrated in Fig. 12.1 *B*.

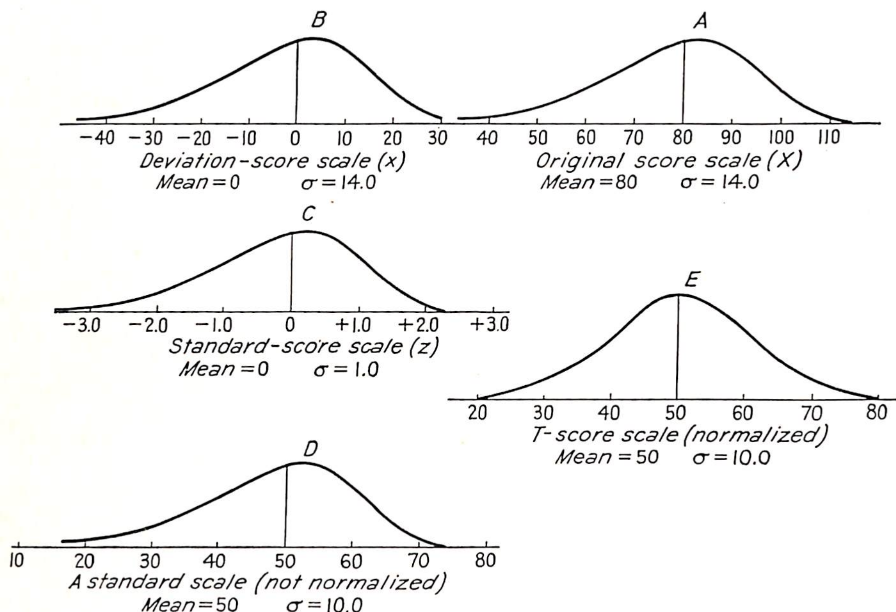


FIG. 12.1.—Distributions before and after conversion from a raw-score scale to a common-score scale with a desired mean and standard deviation, with and without normalizing the distribution.

Deducting the mean from every raw score has the effect of shoving the entire distribution down the same scale so that the mean is zero. The final step, arriving at the z scale, is shown in Fig. 12.1 *C*. Distribution *C* is drawn so that the mean is directly beneath that in distribution *B*, both at zero, and so that deviations of 14 units on the original scale correspond with deviation of 1σ on the standard scale. Especially to be noted is the fact that the form of distribution has not changed; it is still skewed exactly as it was originally. This procedure *does not* normalize the distribution as some other scaling procedures do.

Application to Comparisons of Scores.—The two students represented in Table 12.1 will now be compared in terms of their standard scores. Before

we take these comparisons very seriously, however, we must consider two possible limitations to this procedure. Applying formula (12.1), we arrive at the standard scores in column (6) of Table 12.1. For accurate comparisons between different tests, there are two necessary conditions to be satisfied. The population of students from which the distributions of scores arose must be assumed to have equal means and dispersions in all the abilities measured by the different tests and the form of distribution, in terms of skewness and kurtosis, must be very similar from one ability to another. Unfortunately, we have no ideal scales common to all these tests, measurements which would tell us about these population parameters. Certain selective features might have brought about a higher mean, a narrower dispersion, and a negatively skewed distribution on the actual continuum of ability measured by one test, and a lower mean, a wider dispersion, and a symmetrical distribution on the continuum of another ability represented by another test. Since we can never know definitely about these features for any given population, if we want to achieve communality of scales at all (standard or any other), we often have to proceed on the assumption that actual means, standard deviations, and form of distribution are uniform for all abilities measured. In spite of these limitations, it is almost certain that derived scales, such as the standard-score scale, provide us with more nearly comparable values than do raw-score scales. The recognition of these limitations, however, should be admitted and interpretations based upon the use of standard scores should be made with appropriate reservations in line with those limitations.

Returning to Table 12.1, with the standard scores we have for the two students, we can now give more satisfactory answers to the questions raised above about these students. Student I is most superior in the psychological test, next in scholastic aptitude, and third in English. Had we judged this by his deviations from the mean, we should have decided that his order of superiority was scholastic aptitude first, English second, and psychological third. We find that in terms of standard scores he is equally deficient in reading ability and information, whereas the deviations would have placed him lower in information than in reading. Student II's five standard scores come in about the same rank order as do his deviation scores but certainly not in the same order as his raw scores.

When comparing the two students in terms of raw scores, we should conclude that student I has the greatest advantage in number of points in scholastic aptitude; in terms of deviations, this would be the same, but in terms of standard scores it is in the psychological test that the advantage is greatest. Student II has about the same superiority over student I in the reading and information tests in terms of raw scores

and deviations but has decidedly greater superiority in reading ability in terms of standard scores. When we compare the two students as to total or average score, whereas the raw-score total gives student I the distinct advantage of 37 points, or an *average* superiority of about 7 points, the standard-score averages reverse the order and give student II a 0.39σ lead. In a scholarship contest, we should conclude that student II has the greater all-round ability as indicated by these tests, when students are compared on a standard-score basis.

A Measure of Individual Intravariability.—Studies of variability within persons (intravariability) have often resorted to the use of standard scores. In terms of them, is student I more or less variable than student II? Here the average deviation is a suitable mode of comparison. In column (7) of Table 12.1 is given the absolute deviation of each standard score from the student's own mean score. The average deviations of these two students in the five tests are 1.74 and 1.02, respectively. In other words, student I is about 70 per cent more variable than student II. Although this is the usual procedure for determining intravariability, a word of caution is important. In using this procedure, we are assuming that in all the abilities measured the true variability of the group measured is the same. The standard-score scale makes the distributions all alike, with standard deviations equal to 1.0. Should we happen to have sampled a group that is actually more variable in one ability than in another, we do not really have comparable units of measurement in all tests. This procedure also assumes tests of approximately equal reliability.

Disadvantages of Standard Scores.—Although standard scores will do for us all that we have said and more, under the proper conditions, there are several things about them which make them less convenient than some others. One shortcoming is the fact that half the scores will be negative in sign, which makes things awkward in computation. Another disadvantage is the very large unit, which is one standard deviation.

We could, of course, overcome the first shortcoming by adding a constant to all the scores to make them all positive, and we could multiply them by another constant, preferably by 10, to make the unit smaller and the range in total units greater. If we did both of these we could achieve almost any mean and standard deviation we wanted depending upon the choice of constants. If we wanted a mean of 50 and a standard deviation of 10, we would multiply every standard score by 10 and add 50.

Direct Scaling to a Desired Mean and Standard Deviation.—This brings us to a more general procedure. If we knew from the time that we had acquired the distribution of raw scores that we were to convert them to a common scale with a certain mean and standard deviation, we would not

go to the trouble of converting first to standard scores then to the new scale. We can do the operation in one step by the equation[1]

$$X_s = \left(\frac{\sigma_s}{\sigma_o}\right) X_o - \left[\left(\frac{\sigma_s}{\sigma_o}\right) M_o - M_s\right] \quad \text{(Conversion of scores in one scale directly to comparable scores in another scale)} \quad (12.2)$$

where X_s = a score on the standard scale, corresponding to X_o.
X_o = a score on the obtained scale; a raw score.
M_o and M_s = means of X_o and X_s, respectively.
σ_o and σ_s = standard deviations of X_o and X_s, respectively.

If the desired mean is 50 and the desired standard deviation is 10, with these substitutions the equation becomes

$$X_s = \left(\frac{10}{\sigma_o}\right) X_o - \left[\left(\frac{10}{\sigma_o}\right) M_o - 50\right]$$

Knowing σ_o and M_o from the particular distribution of raw scores, the equation reduces to very simple form describing a straight line. Taking the illustration of Fig. 12.1, where $M_o = 80$ and $\sigma_o = 14.0$,

$$X_s = \left(\frac{10}{14}\right) X_o - \left[\left(\frac{10}{14}\right) 80 - 50\right]$$
$$= .714X_o - 7.12$$

A raw score of 100 would, by this formula, become a scaled score of 64. A raw score of 50 would become a scaled score of 29. We can see a graphic exhibition of this transformation by relating distributions A and D in Fig. 12.1. A score of 100 in A is in a position comparable to a score of 64 in D, and a score of 50 in A is in position similar to 29 in D.[2]

Scaling by this procedure, as by the standard-score method, assumes that the obtained form of distribution is the same as the population distribution. If this is true, then it is probable that units on the derived scale are equal, also those on the raw-score scale. So far as improving the equality of units is concerned, then, nothing has been gained, nor was anything to be gained. We know, however, that the form of distribution of a sample is not necessarily the form of distribution of the population. The discrepancy need not be, and probably is not, due to sampling errors, particularly if the sample is large. There are many reasons for radical departures of sample distributions from genuine population distributions of the trait measured:

[1] For the derivation of this type of equation, see Appendix A.

[2] A more general discussion of such a transformation procedure will be found in Ch. 19. Formula (12.2) here is the same as formula (19.8) there, with a change in symbols.

difficulty level of the test, intercorrelation of the items (see Ch. 17), and the variations in difficulty and intercorrelation. We should not, therefore, feel too obligated to retain the same form of distribution in scaled scores as in the raw scores. If there is a real discrepancy between population distribution and sample distribution, there is much room for improvement of the scale in terms of equality of units. The next methods to be described have the probable advantage that by normalizing distributions they also achieve better metric scales.

The *T* Scale and *T* Scaling of Tests

The well-known *T* scale overcomes the objections raised against standard scores and adds besides an advantage peculiar to itself. It adopts as its unit one-tenth of a standard deviation, so that an ordinary distribution with a range of 5 to 6σ on its base line yields 50 to 60 integral *T*-scale scores. In addition, the *T* scale goes beyond any ordinary dis-

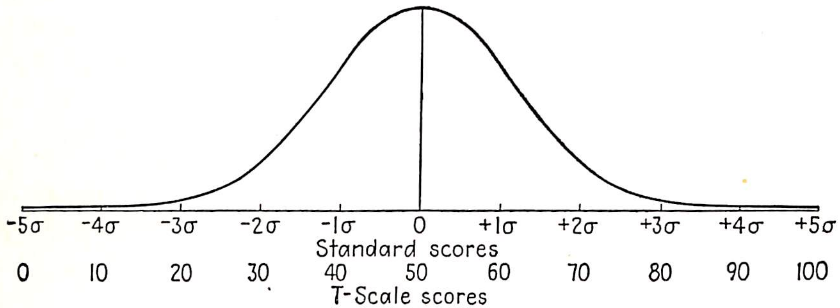


Fig. 12.2.—The *T* scale and its relation to the standard-score scale extending over a range of 10 sigma units.

tribution, extending its scale over a spread of 10 standard deviations, or 100 units in all. Any age or grade group would yield its own distribution extending 5 to 6σ. A group just higher in ability would overlap this one and yet would need an extension over new units beyond the limit of the first group. A third group of lower age would need an extension of the measuring stick at the other end. When all groups from lowest to highest are taken into account, considerable extension is required. The result, with these extensions, is a single common scale on which all groups, over a wide range, have a common unit and a common zero point. It has been found in practice that a scale with 100 units (or 10σ) will be extensive enough. It is based upon a normal curve whose tails extend from -5σ to $+5\sigma$ (see Fig. 12.2). Besides making the unit equal to 0.1σ, the *T* scale also has the zero point at the extreme left, which places it at -5σ. The mean now becomes 50, and the other *T*-scale points are spaced as in Fig. 12.2.

McCall, who originated the *T* scale, suggested that the mean of this curve should be that of a representative twelve-year-old group. This mean was chosen because the twelve-year-olds are about midway along the scale of development. Since any limited sample of them would range over not more than about 60 units of the *T* scale, groups of higher and lower ability were required to complete the picture and to determine what kind of performance comes at 80 to 100 points at the upper end of the scale and 0 to 20 at the lower end. The method of finding *T*-scale equivalents for performances beyond the ranges of tested samples will not be described here. Suffice it to say that many test makers take pains to set up means of converting raw scores on all tests into *T*-score equivalents. The *T*-scale principle can be used with any standard group of individuals, whether they are twelve-year-olds or not. The procedure for converting raw scores in any test into *T*-scale equivalents (though not with the twelve-year-old mean and unit) will now be described.

How to Derive *T*-scale Equivalents for Raw Scores.—A college or university or a single school system may wish to use the *T*-scale idea as its common yardstick for all its tests. The freshmen entering a large university, for example, may be taken as the standard group for this purpose. As an illustration, let us use the data in Table 12.2. Here is a distribution of 83 scores obtained by freshmen in an English examination of the objectively scored type. The procedure will be described step by step:

Step 1. List the class intervals as usual. Here a maximum number of class intervals is best; 20 or even more.

Step 2. List the exact upper limits of class intervals.

Step 3. List the frequencies.

Step 4. List the cumulative frequencies (see Ch. 6 for instructions).

Step 5. Find the cumulative proportions for the class intervals.

Step 6. Find the corresponding *T* scores from Table 12.3. These are then listed in the last column of Table 12.2, given to one decimal place. We usually want finally a ready means of reading directly the *T* score corresponding to any integral raw score. It is recommended that the remaining steps be taken to satisfy this objective.

Step 7. Plot a series of points to represent each *T* score in Table 12.2 corresponding to the upper limit of the class interval, as in Fig. 12.3. If the original distribution of raw scores is normal, the points should fall rather close to a straight line. The reason that they are not perfectly in line is that there are some irregularities in the original data. Draw through the points with a straight edge a line

Table 12.2.—The Calculation of T scores for a Distribution of English-examination Scores

(1) Scores	(2) Upper limit of interval	(3) Frequency	(4) Cumulative frequency	(5) Cumulative proportion	(6) *T* score (from Table 12.3)
225–229	229.5	1	83	1.000	—
220–224	224.5	0	82	.988	72.6
215–219	219.5	1	82	.988	72.6
210–214	214.5	5	81	.976	69.8
205–209	209.5	5	76	.916	63.8
200–204	204.5	7	71	.855	60.6
195–199	199.5	6	64	.771	57.4
190–194	194.5	6	58	.700	55.2
185–189	189.5	6	52	.627	53.2
180–184	184.5	11	46	.554	51.4
175–179	179.5	9	35	.422	48.0
170–174	174.5	5	26	.313	45.1
165–169	169.5	5	21	.253	43.3
160–164	164.5	6	16	.193	41.3
155–159	159.4	5	10	.120	38.2
150–154	154.5	2	5	.060	34.5
145–149	149.5	1	3	.036	32.0
140–144	144.5	1	2	.024	30.2
135–139	139.5	0	1	.012	27.4
130–134	134.5	1	1	.012	27.4

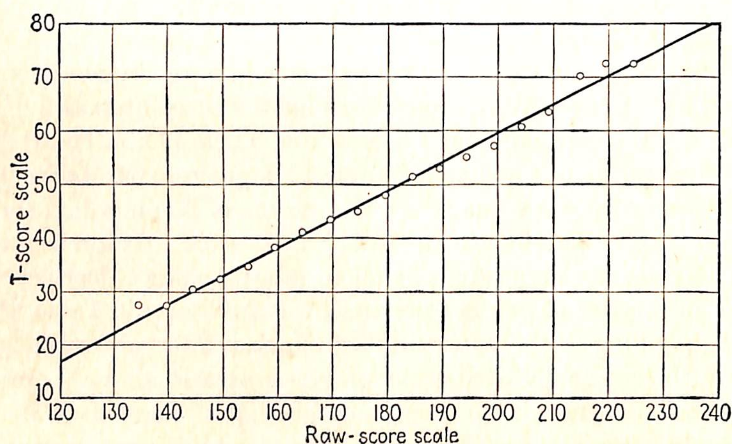


Fig. 12.3.—A smoothing process applied in deriving *T*-scale equivalents for English-examination scores (see Table 12.2).

TABLE 12.3.—A TABLE TO AID IN THE CALCULATION OF T SCORES

Proportion below the Point	*T* score	Proportion below the Point	*T* score	Proportion below the Point	*T* score
.0005	17.1	.100	37.2	.900	62.8
.0007	18.1	.120	38.3	.910	63.4
.0010	19.1	.140	39.2	.920	64.1
.0015	20.3	.160	40.1	.930	64.8
.0020	21.2	.180	40.8	.940	65.5
.0025	21.9	.200	41.6	.950	66.4
.0030	22.5	.220	42.3	.960	67.5
.0040	23.5	.250	43.3	.965	68.1
.0050	24.2	.300	44.8	.970	68.8
.0070	25.4	.350	46.1	.975	69.6
.010	26.7	.400	47.5	.980	70.5
.015	28.3	.450	48.7	.985	71.7
.020	29.5	.500	50.0	.990	73.3
.025	30.4	.550	51.3	.993	74.6
.030	31.2	.600	52.5	.995	75.8
.035	31.9	.650	53.9	.9960	76.5
.040	32.5	.700	55.2	.9970	77.5
.050	33.6	.750	56.7	.9975	78.1
.060	34.5	.780	57.7	.9980	78.8
.070	35.2	.800	58.4	.9985	79.7
.080	35.9	.820	59.2	.9990	80.9
.090	36.6	.840	59.9	.9993	81.9
		.860	60.8	.9995	82.9
		.880	61.7		

that will come as close to all the points as seems possible. Among those that do not touch the line, as many of them should be above it as below it. The line may be extended beyond the ends of the points at both ends. If the raw-score distribution is skewed, the trend in the points when plotted will show some curvature. It is best, then, to attempt to follow the curvature but with a smooth trend. If the curvature is not followed, the distribution of the population on the scaled scores will not be normalized.

Step 8. For any integral raw-score point, we can now find the corresponding *T*-score points. For example, in Fig. 12.3, a raw score of 220 corresponds to a *T* score of 70, and a raw score of 150 corresponds to a *T* score of 33. In this we favor integral *T* scores but

at times have to resort to half points when we cannot decide upon the nearest unit.

Step 9. Prepare a table in which every integral raw score, or every second, third, or fifth one, appears in one column and the corresponding T scores in the other. Table 12.4 is such a tabulation. It will serve for all future purposes of translation where the original tested group remains the standard. Many test users prefer to list *every* raw score and its T-score equivalent so as to avoid the need for interpolation.

TABLE 12.4.—RECTIFIED SCALING WITH T SCORES FOR THE DISTRIBUTION OF ENGLISH-EXAMINATION SCORES

Examination score	T score	Examination score	T score	Examination score	T score
240	81	195	57	155	35.5
235	78	190	54	150	33
230	75.5	185	51.5	145	30
225	73	180	49	140	27.5
220	70	175	46	135	25
215	67.5	170	43.5	130	22
210	65	165	41	125	20.5
205	62	160	38	120	17
200	59.5				

A Normal Graphic Procedure for T Scaling.—It is possible to do more of the T scaling graphically by the use of normal probability paper. This graph paper is especially designed with spacing for cumulative proportions along one axis in a manner consistent with the cumulative normal-curve function. Figure 12.4 shows how the English examination data can be so treated. Using the cumulative proportions appearing in Table 12.2, column (5), we plot each one against its corresponding raw-score value given in column (2). The trend of the points will be in a straight line if the distribution of raw scores is normal. If that distribution is skewed there will be some curvature in the trend which one should try to follow in smoothing. To find the T equivalent for any raw score, we find that raw score on the base line, follow it up to the line drawn through the points, locate the equivalent proportion, then go to Table 12.3 for the corresponding T.

An Evaluation of the T-scale Procedure.—The T scale is probably the most widely used of all derived scales. Its advantages are many, its disadvantages few. When the scaling is carried out, as described, the

procedure normalizes distributions. This effect is pictured in Fig. 12.1. Contrast distributions *D* and *E* in that illustration. Both have a mean of 50 and a σ of 10. The one is skewed like the original distribution, the other is normal. The normalizing process comes about through the conversion to centiles and then to corresponding deviations from the mean in a normal distribution. Table 12.3 is based upon the normal curve. For a given proportion (area below a given point) is given a *T*-score equivalent instead of a standard-score equivalent.

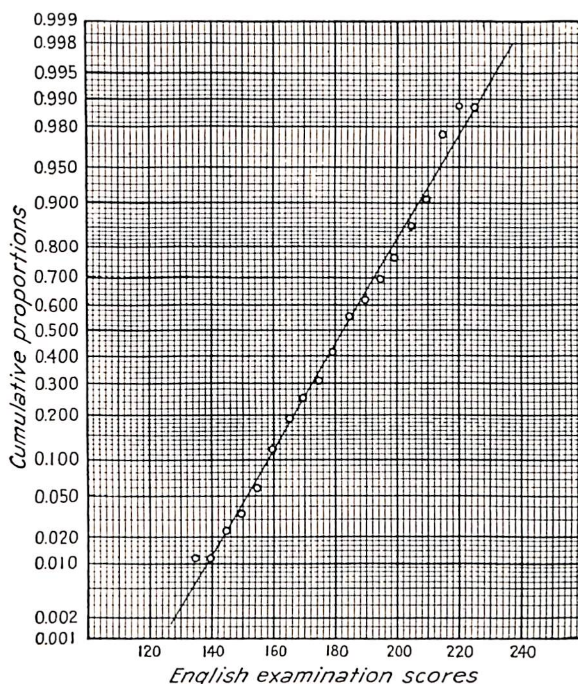


FIG. 12.4.—A graphic solution to scaling which utilizes normal probability graph paper.

The normalizing process may be pictured as in Fig. 12.5. There the obtained distribution, seriously skewed, is given below, and the normalized distribution on the derived scale above. The process assures that the *areas* *A*, *B*, *C*, . . . , *M* correspond in the proportions that they occupy with *areas* *A′*, *B′*, *C′*, . . . , *M′*. The correspondences of scale distances are also shown, by connecting dotted lines. If the units on the derived scale (not shown) represent genuinely equal increments of the measured variable, then obviously those on the original scale do not. We may not know that the population is normally distributed on a trait, but by normalizing distributions, where there is no inhibiting information to the contrary, we achieve more common and meaningful scores.

Other advantages of the *T* scale have been mentioned—the possibility of extending it beyond limited populations, its convenient mean, unit, and standard deviation, and its general applicability. It has some limitations which should be pointed out. In much practical use of tests, as fine a unit as $.1\sigma$ may be an overrefinement. Much coarser discriminations are all that may be necessary. Furthermore, the unit may give quite a false sense of accuracy of the measurement that is actually being made. If the original scores had a standard deviation much smaller than 10—for example, one of 5 score units—then the substitution of a unit of $.1\ \sigma$ is in a sense "hairsplitting." Two whole units on the *T* scale are then as fine a distinction as we could actually make between individuals. Nor is this the

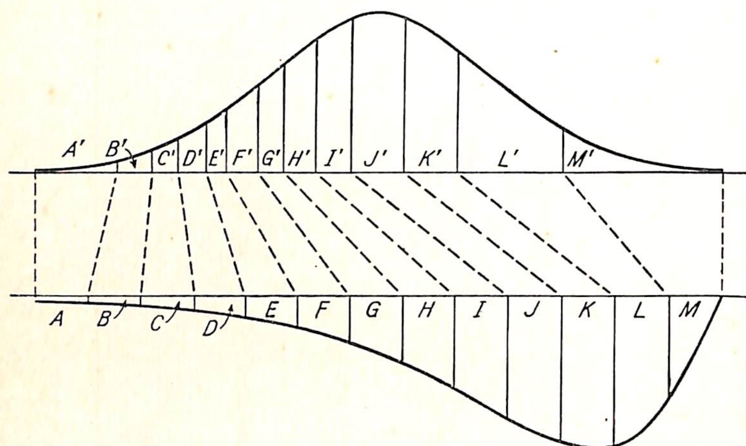


Fig. 12.5.—A graphic illustration of what happens in scaling so as to normalize a distribution. Intervals are matched so as to equate corresponding *areas* under the curves.

whole story. Every test, even the best of them, has an error of measurement whose size is indicated by its "standard error of an obtained score" (see Ch. 17). This stems from the fact that the test is not perfectly reliable. If the error of measurement is as much as 2 units on the raw-score scale, it might be even larger on the *T* scale. If the error is such that the best practical discriminations we can make between individuals is of the order of one-half σ, it is rather presumptuous to apply a scale that pretends to distinguish to one-tenth σ. For this reason, particularly, and because many test users require less refinement than the *T* scale offers, the writer has proposed the *C* scale, which will be described next.

The *C* Scale and *C* Scaling

The *C*-scale System.—The principles of the *C* scale and the derivation of *C*-scale equivalents for raw scores are illustrated in Table 12.5. The *C*

TABLE 12.5.—THE ELEVEN-POINT SCALED-SCORE SYSTEM AND ITS APPLICATION TO THE MEMORY-TEST DATA

(1) *C*-scale score	(2) Standard-score limits	(3) Centile-rank limits	(4) Percentage within each interval	(5) Percentage in whole numbers	(6) Corresponding score points in the memory test	(7) Memory-test scores in each scaled-score interval
	+2.75	99.7				
10			0.9	1		41+
	+2.25	98.8			40.5	
9			2.8	3		38–40
	+1.75	96.0			37.6	
8			6.6	7		35–37
	+1.25	89.4			34.6	
7			12.1	12		31–34
	+0.75	77.3			30.8	
6			17.4	17		28–30
	+0.25	59.9			27.8	
5			19.8	20		25–27
	−0.25	40.1			24.4	
4			17.4	17		21–24
	−0.75	22.7			20.8	
3			12.1	12		18–20
	−1.25	10.6			17.7	
2			6.6	7		15–17
	−1.75	4.0			14.5	
1			2.8	3		12–14
	−2.25	1.2			11.8	
0			0.9	1		0–11
	−2.75	0.3				

scale is so arranged that the mean will be exactly at 5.0, with the two limiting classes being 0 and 10. Column (2) gives the exact limits of the 11 units in terms of standard scores. The corresponding centile limits (derived from Table B) are given in column (3) The percentage of cases within each unit is found by subtracting neighboring pairs of centile limits. Thus, in the middle unit, the difference 59.9 − 40.1 = 19.8, etc. Since it is more convenient to think in terms of whole numbers, the approximate percentages of the cases falling in the different classes are given as nearest whole numbers in column (5). These can be used either as a guide in thinking of the make-up of the standard distribution or even in subdividing lists of scores of individuals when arranged in rank order. Thus,

if we had 100 persons lined up in rank order in a test, the highest person would be given the score of 10, the next 3 a score of 9, the next 7 a score of 8, etc., until the last in line is given a score of 0.

Steps in Deriving a *C* Scale.—The operations for deriving a *C* scale are much the same as those for deriving a *T* scale. There are some differences in the steps to be recommended, however, so all the steps will be listed here.

Step 1. List the class intervals.
Step 2. List the exact upper limits of the intervals.
Step 3. List the frequencies.
Step 4. List the cumulative frequencies.
Step 5. Find the cumulative proportions for the intervals.
Step 6. From here on the steps differ from those for *T* scaling. Next, plot the cumulative proportions on the ordinate corresponding to *X* values (exact upper limits) on the abscissa of coordinate paper. (See Ch. 6 for further instructions.)
Step 7. Draw by inspection a smooth S-shaped curve through the trend of the points. If the distribution is obviously skewed and one tail of the S is short, or even if it vanishes, follow the points anyway. At this stage one sees the advantage of having a liberal number of classes.
Step 8. Look for each of the centile limits [from column (3) of Table 12.5] on the ordinate, find the intersection of that centile-rank level with the curve, drop down to the abscissa to locate the corresponding raw-score point. Try to avoid arriving at a point exactly at integers, so that it is clear whether each integral raw score goes above or below the division point. The values obtained from this step are like those in column (6) of Table 12.5.
Step 9. Determine within which *C* intervals the various integral score values lie and write the limiting scores as in column (7) of Table 12.5.

Alternative Graphic C-scaling Steps.—If one already has a figure drawn like Fig. 12.3 that is used in *T* scaling, one could use it to accomplish steps 6 and 7 in the following manner. The σ for the *T* scale is 10 and that for the *C* scale is 2. The means are 50 and 5, respectively. An interval of one unit on the *C* scale corresponds to five units on the *T* scale. A *C* score of 5, therefore, occupies a range from 47.5 to 52.5; a *C* score of 6 corresponds to a range 57.5 to 62.5, and so on. All the *T*-score limits of the *C* intervals can be seen represented in Table 12.6. The *T*-score limits, therefore, can be

TABLE 12.6.—*T* SCORES EQUIVALENT TO *C*-SCORE INTERVALS

C score	*T*-score limits	Middle *T* score
10	72.5–77.5	75
9	67.5–72.5	70
8	62.5–67.5	65
7	57.5–62.5	60
6	52.5–57.5	55
5	47.5–52.5	50
4	42.5–47.5	45
3	37.5–42.5	40
2	32.5–37.5	35
1	27.5–32.5	30
0	22.5–27.5	25

located in Fig. 12.3 and from them the corresponding points of division on the raw-score scale. These mark off the raw-score ranges corresponding to all *C* scores.

The normal-graphic procedure described in connection with *T* scaling can also be applied here; in fact, it is even more convenient in this connection and is to be recommended in preference to steps 6 and 7. Since the centile ranks are marked on probability paper (see Fig. 12.4), one would locate the centile-rank limits [column (3) of Table 12.5] and from the plot, usually a straight line, find the corresponding raw-score division points.

An Evaluation of the *C* Scale.—The *C* scale has many of the advantages of the *T* scale. It refers obtained scores to a common scale that is related to the normal distribution. If the population distribution on a measured trait is normal, then the distribution of *C* scores properly represents that population and the units of measurement may be regarded as equal. It lacks the refinement of a small unit such as that provided by the *T* scale. On the other hand, it probably more nearly represents the accuracy of discrimination actually made by means of tests, and its broader categories will do for guidance purposes. There is a handicap in selection of personnel in that a change of minimum qualifying score of only one *C*-scale unit may result in quite a difference in percentage of cases selected. For example, if the cutoff score were changed from 5 to 6, 20 per cent more rejections would have to be made. For selection purposes, however, raw-score cutoffs may be just as feasible as derived scores. The reference of any chosen raw-score cutoff to equivalent *C*-score limits or centiles would add meaning to that particular value.

For guidance and counseling purposes, the use of a zero *C* score may be unwise. Unless he is more sophisticated than most people, a counselee would hardly relish being told that he earned a score of zero. To meet this contingency, one could let the scores range from 1 through 11 instead of 0 through 10. Or one could resort to a condensed scale to be described next.

The Stanine Scale.—There are several reasons for condensing the *C* scale to some extent by giving it a 9-unit range. This is usually done by combining the two categories at either end, with 4 per cent of the distribution in categories 1 and 9. Such a scale was standard for the Army Air Forces Aviation Psychology Program during World War II. All test scores and composites were eventually scaled to this system, called "stanine" as a contraction of "standard nine." The mean of such a norm distribution would be 5.0, as in the *C* scale, but the standard deviation would be slightly lower—1.96—because of the contractions at the tails of the curve.

Perhaps the chief practical benefit to be derived from 9 units rather than 11 is that such scores occupy only one column on the IBM punched-card records. For research purposes, however, a significant grouping error (see Ch. 5) is thus introduced, calling for corrections of various sorts when precise statistics are wanted. In guidance work, many counselors would probably not like to have the rare one person in a hundred at either extreme submerged with the other three per cent next to him. There is probably a full unit's discrimination between the hundredth person and the next three per cent just as there is between any other neighboring categories. This loss of discrimination in the stanine scale may not be tolerated and is unnecessary in the use of profiles in guidance.

Some Norm and Profile Suggestions

Suggestions were made in Ch. 6 concerning the derivation of centile norms and the construction of profiles. Here we are ready for other, more comprehensive, suggestions. There will be shown both a norm table and a profile chart, in each of which raw scores can be interpreted in terms of the *C* scale, *T* scale, or centile rank.

Conversion Table for Deriving Scaled Scores and Centiles.—Table 12.7 provides an example of a norm table based upon five parts of the *Guilford-Zimmerman Aptitude Survey*. The five parts represented include tests of *Verbal Comprehension*, *General Reasoning*, *Numerical Operations*, *Perceptual Speed*, and *Spatial Orientation*, respectively. By means of the table any raw score can be readily transformed into a corresponding common-scale score. A raw score of 45 in Part I, for example, means a *C* score of 6; by interpolation, a centile rank of 68; and a *T* score of 55. Each test user

TABLE 12.7.—EXAMPLE OF RECOMMENDED GENERAL-PURPOSE NORMS, BASED UPON FIVE PARTS OF THE GUILFORD-ZIMMERMAN APTITUDE SURVEY TESTS (FORM *A*)

C score	Part I	Part II	Part III	Part IV	Part V	Centile limits	*T* score
10	68+	25+	110+	70+	52+	99+	75.0
							72.5
9	62–67	23–24	95–109	66–69	46–51	96–98	70.0
							67.5
8	56–61	21–22	86–94	60–65	39–45	89–95	65.0
							62.5
7	49–55	18–20	77–85	54–59	33–38	77–88	60.0
							57.5
6	42–48	15–17	68–76	49–53	28–32	60–76	55.0
							52.5
5	34–41	12–14	59–67	44–48	22–27	40–59	50.0
							47.5
4	27–33	9–11	50–58	39–43	16–21	23–39	45.0
							42.5
3	21–26	6–8	41–49	33–38	10–15	11–22	40.0
							37.5
2	17–20	4–5	32–40	27–32	5–9	4–10	35.0
							32.5
1	13–16	2–3	23–31	21–26	2–4	1–3	30.0
							27.5
0	< 13	< 2	< 23	< 21	< 2	< 1	25.0

would probably choose his own mode of interpretation and make only one of these conversions for each score. The most ready interpretation possible here is to the *C* scale. If one had many conversions to make to centile ranks, or to *T* scores, one would probably prefer a table requiring less interpolation. Such a table would need more intervals than 11, or different limits, for greater convenience.

A Profile Chart with Three Interpretive Scales.—Fig. 12.6 shows a profile chart that can be used to provide not only a general contour representing several traits for an individual, but also reference to three common scales. The basic intervals are, again, the *C*-score units. Corresponding to each *C* score are given for each test the two limiting raw scores for that interval. At the lower margin are provided the single *T* score and the centile rank corresponding to the midpoint of each interval. Finer decisions on both these scales can be made, if desired, by interpolation.

The illustration shows the record for a certain individual who earned raw scores of 28, 88, 21, and 23 in the Memory, Vocabulary, Word Building, and Sentence Construction tests, respectively. From the chart the *C* scores

are obviously 6, 7, 6, and 4, respectively. The nearest centile ranks are 70, 84, 70, and 30, respectively. If we want more exact centile-rank estimates, by interpolation we get something like 67, 82, 73, and 25. It is doubtful, however, whether the actual refinement of the tests justifies distinctions as fine as these. There is little virtue in quibbling over a difference of 2 to 5 centile rank points (except at the extremes of the range) when the error of measurement may be as much as twice as large. In terms of T scores, the nearest values are 55, 60, 55, and 45, respectively. Again, finer estimates could be made but are probably not necessary or entirely justifiable. It may be sufficient, for guidance purposes, at least, to note that the individual represented is on the under side of a C score of 6

	0	1	2	3	4	5	6	7	8	9	10	*C score*
Memory	11	14 12	17 15	20 16	24 21	27 25	30 28	34 31	37 35	40 38	41+	
Vocabulary	56	62 57	67 63	72 68	77 73	81 78	85 82	89 86	93 90	96 94	97+	
Word Building	6	9 7	11 10	13 12	16 14	19 17	21 20	24 22	27 25	33 28	34+	
Sentence Construction	16	18 17	20 19	22 21	24 23	26 25	28 27	30 29	34 31	38 35	39+	
	1	2	7	16	30	50	70	84	93	98	99	*Centile rank*
	25	30	35	40	45	50	55	60	65	70	75	*T score*

FIG. 12.6.—Example of norms on four tests in terms of C scores, T scores, and centile ranks. A profile has been drawn for one individual.

in the Memory test, just above the middle of the C score of 7 for the Vocabulary test, in the upper half of the C score of 6, and on the lower side of the C score of 4 for World Building, if not to be content with the simple results of 6, 7, 6, and 4.

Exercises

1. Determine the standard scores for the two students in Data 12*A*. Draw conclusions as to how the two students compare with respect to raw scores and with respect to standard scores. Which is probably the better student in terms of aptitude (assume that all five tests measure different aspects of aptitude about equally well)? Which is the more variable student? Support your answers with evidence.

2. Derive a conversion equation for translating scores in the Syllogism test into a scale which would give a mean of 50 and a standard deviation of 10. Using the equation, what would the scores for students A and B become on the new scale?

3. Determine the equivalent T scores for one or more of the distributions in Data 12*B*.

4. By a graphic process find by smoothing a modified set of equivalent T scores for the same distribution or distributions.

5. Using the T-score equivalents found in Exercise 3 or Exercise 4, convert the raw scores for students A to D in Data 12*C* into corresponding T scores.

6. Determine the score limits for one or more of the tests in Data 12*B* corresponding to each *C*-score category. Use a smoothing process applied either to the ogive or to the line drawn on probability paper.

7. Prepare a profile chart based on *C*-score units (but including also equivalent *T* scores and centiles) for the three tests in Data 12*B*.

8. Report equivalent *C* scores for the four students in Data 12*C*.

DATA 12*A*.—MEANS AND STANDARD DEVIATIONS IN FIVE PARTS OF AN ENGINEERING-APTITUDE EXAMINATION (ROUNDED TO WHOLE NUMBERS) AND SCORES OF TWO STUDENTS

Test	Figure classification	Cube visualizing	Syllogism	Paper folding	Form perception
Mean	22	15	28	33	26
σ	4	6	8	5	7
Student *A*	28	26	30	17	35
Student *B*	15	32	15	32	41

DATA 12*B*.—FREQUENCY DISTRIBUTIONS OF ENGINEERING FRESHMEN IN THREE APTITUDE TESTS

Scores	Cube visualizing	Syllogism	Form perception
45–49	...	4	...
40–44	...	13	2
35–39	...	29	16
30–34	1	42	42
25–29	8	45	52
20–24	35	43	55
15–19	58	24	26
10–14	63	6	13
5– 9	36	...	1
0– 4	6	...	...
Sums	207	206	207

DATA 12*C*.—SCORES OF FOUR STUDENTS IN THE THREE TESTS REPRESENTED IN DATA 12*B*

Student	Cube	Syllogism	Form
A	25	22	34
B	5	45	12
C	33	42	37
D	11	20	16

CHAPTER 13

SPECIAL CORRELATION METHODS AND PROBLEMS

Pearson's product-moment coefficient is the standard index of the amount of correlation between two things, and we employ it whenever it is possible and convenient to do so. But there are data to which this kind of correlation method cannot be applied, and there are instances in which it can be applied but in which, for practical purposes, other procedures are more expedient. The Pearson coefficient cannot or should not be computed, for example, unless the two variables X and Y are measured on continuous metric scales and unless the regressions are linear (see Ch. 15). Many of our data are in terms of frequencies of cases having attributes; they are on variables of a "qualitative" rather than a quantitative sort. Less often, two continuously measured variables bear to one another a relationship that is curved rather than in the form of a straight line. In this chapter will be described some procedures that take care of these irregular situations and of other situations where short-cut methods are better used to compute a Pearson r or its equivalent.

Even when we can apply the product-moment correlation method, however, there are many circumstances which may give rise to a somewhat different estimate of correlation than is typical or to one that does not apply to the population in which we are interested. Samples may be heterogeneous or they may be restricted in variability or they may be forced into a smaller number of categories than we need for good estimates of correlation, free from errors of grouping. These, and other common irregularities in the sampling situation or in the data, call for special corrective steps and for special interpretive action. It is impossible to anticipate all the peculiarities of data that the reader may encounter, but the more common exceptions to ideal correlation conditions will be touched upon.

Spearman's Rank-difference Correlation Method

When samples are small, a common procedure applied to regular data in the place of the product-moment method is the rank-difference method of Spearman. It is conveniently applied as a quick substitute when the number of pairs, or N, is less than 30. It is even more conveniently applied

when the data are already in terms of rank orders rather than in terms of measurements.

The Computation of a Spearman Rho.—If we have data in terms of measurements or scores, it is first necessary to translate them into rank orders. The procedure will be demonstrated by means of the data in Table 13.1. There we have 15 pairs of scores for 15 individuals who

TABLE 13.1.—A RANK-DIFFERENCE CORRELATION BETWEEN HUMOR SCORES IN REACTIONS TO CARTOONS AND TO LIMERICKS

Cartoon score	Limerick score	R_1	R_2	D	D^2
47	75	11	8	3	9.00
71	79	4	6	2	4.00
52	85	9	5	4	16.00
48	50	10	14	4	16.00
35	49	14.5	15	0.5	0.25
35	59	14.5	15	0.5	0.25
41	75	12.5	8	4.5	20.25
82	91	1	3	2	4.00
72	102	3	1	2	4.00
56	87	7	4	3	9.00
59	70	6	10	4	16.00
73	92	2	2	0	0.00
60	54	5	13	8	64.00
55	75	8	8	0	0.00
41	68	12.5	11	1.5	2.25
					165.00 ΣD^2

responded to sets of cartoons and limericks by judging their humor values, each on a 5-point scale. The score in each case is the sum of the points each individual assigned to the set. We could correlate these scores in the usual manner, described in Ch. 8. The rank-difference method will be found shorter. The following steps are necessary:

Step 1. Rank the individuals from the highest to the lowest in the first variable (here it is "cartoon score"), and call these ranks R_1. The highest score receives the rank of 1 (which is arbitrary; we might have called it 15), the next highest 2, etc. The only difficulty encountered is when we find ties. For example, in Table 13.1, two individuals have scores of 41. One of them comes at

rank 12 and the other at rank 13. We do not know which, if either, is better, yet we must fill these two rank positions; so we take the average of the tied ranks and call them both 12.5. We make certain that the next ranking scorer is called 14, unless he also is tied. We find that he is tied with another who has a score of 35. We treat these two in a similar manner; so they become each 14.5. If the lowest person is not tied with others, the last rank should be equal to N (in this case, 15). This serves as a check as to accuracy of ranking, though, of course, it will not detect inversions in rank order somewhere along the line. It merely shows whether any rank has been repeated, whether any individuals have been overlooked, or whether ties have somewhere not been properly treated.

Step 2. Rank the second list of measurements in a similar manner, and call them R_2. In this problem, there are three scores of 75 for the individuals occupying places 7, 8, and 9. We call them all 8, leaving out of the list 7 and 9. This treats the three alike, as they should be, yet gives us a full set of 15 ranks.

Step 3. For every pair of ranks (for each individual), determine the difference in ranks. The smaller one can be subtracted from the larger one in each case, with no attention being paid to algebraic signs, for they are all going to be squared anyway.

Step 4. Square each difference to find D^2.

Step 5. Sum the squares of the differences (see the last column of Table 13.1) to find ΣD^2. The sum in our illustrative problem is 165.00.

Step 6. Compute the coefficient ρ (Greek letter rho) by means of the formula

$$\rho = 1 - \frac{6\Sigma D^2}{N(N^2 - 1)} \quad \text{(Rank-difference coefficient of correlation)} \quad (13.1)$$

where ΣD^2 = sum of the squared differences between ranks.
N = number of pairs of measurements.

In this problem

$$\begin{aligned}\rho &= 1 - \frac{6 \times 165}{15 \times 224} \\ &= 1 - .295 \\ &= .705, \text{ or } .70\end{aligned}$$

By this procedure, then, the estimate of the amount of correlation between the two sets of scores is .70. How shall we interpret this correlation, as compared with a Pearson r?

Interpretation of a Rho Coefficient.—The rank-difference coefficient is practically equivalent to the Pearson r numerically. There is a conversion formula by which the corresponding Pearson r can be estimated from rho. But this formula assumes large samples, which is precisely what we do not have when we compute rho, and in no case is the difference between rho and r greater than .018, and in every case, except for coefficients of zero or 1.00, r is greater than rho. We may therefore treat an obtained rho as an approximation to r and under these circumstances interpret the outcome of a correlation study accordingly.

The estimation of the reliability of rho, as indicated by its standard error, is in some doubt, but a rough approximation is given by the formula[1]

$$\sigma_\rho = \frac{1.04(1 - \rho^2)}{\sqrt{N - 1}} \qquad (13.2)$$

For the illustrative problem,

$$\sigma_\rho = \frac{1.04(1 - .497025)}{\sqrt{14}}$$
$$= .14$$

Formula (13.2) indicates that the sampling error for ρ is only about 4 per cent greater than for a corresponding Pearson r. The obtained σ_ρ of .14 indicates that the obtained ρ of .70 is not a very accurate estimation of the correlation in the population. The standard error of .14 would allow considerable dispersion in sample ρ's for samples of this size. Can we feel confident that the obtained coefficient indicates any positive correlation at all? Reference to Table D in which we look for 13 degrees of freedom shows that it requires a Pearson r of .514 to be significant at the 5 per cent level and an r of .641 to be significant at the 1 per cent level. The similar confidence levels for ρ would have to be about 4 per cent higher, or .54 and .67, respectively. There is less than one chance in 100, apparently, that such a correlation as .70 could have happened if there really were no correlation between the two variables—cartoon scores and limerick scores.

A Method of Dealing with Ties.—DuBois has shown that the procedure of giving tied scores a common rank equal to the mean of the ranks involved in the ties is a good approximation, but that when more than two or three scores are tied, a better estimate is desirable. His formula is

$$R_c = \sqrt{M^2{}_R + \frac{n^2 - 1}{12}} \qquad \text{(Tied rank corrected)} \qquad (13.3)$$

[1] Thornton, G. R. The significance of rank difference coefficients of correlation. *Psychom.*, 1943, **8**, 211–222; Olds, E. G. Distribution of sums of squares of rank differences for small numbers of individuals. *Ann. math. Stat.* 1938, **9**, 133–148.

where M_R = mean of the ranks for the ties.
n = number tied.

Every case of ties would have to be treated separately. For example, in our second set of scores in Table 13.1, there are three ties for eighth place (or three identical scores for places 7, 8, and 9). Applying DuBois's formula,[1]

$$R_c = \sqrt{8^2 + \frac{9-1}{12}}$$
$$= \sqrt{64 + .67}$$
$$= \sqrt{64.67}$$
$$= 8.04$$

Had we used 8.04 instead of 8 in the computation of rho, the result would hardly have been affected. With five or more ranked scores tied, however, the correction would probably make an appreciable difference.

A Brief Evaluation of the Rank-difference Correlation.—Because of its slightly larger standard error than for the Pearson r, and because its use is limited to small samples, it is doubtful whether Spearman's ρ should be used for any purpose except to examine the possibility of correlation between two variables and not to estimate the size of the correlation coefficient unless it happens to be very large (not less than .70), and then only as a rough estimate. Its use for research purposes is therefore very limited. It is a product-moment coefficient and, as such, it approximates the corresponding Pearson r rather closely. The difference between the two is ordinarily well below the size of the standard error of either. There appear to be no specific requirements laid down by writers who describe this coefficient, but since it is a product-moment r and is an estimate of the Pearson r, its use for that purpose presumably rests upon the same assumptions as are necessary in that connection—linear regression and homoscedasticity, in particular.

The Correlation Ratio

The correlation ratio is a very general index of correlation particularly adapted to data in which a curved regression prevails. Among test scores, linear relationships are apparently the almost universal type of regression. Normality, or near normality, in both distributions correlated is almost sufficient in itself to promote linearity. Outside the sphere of psychological and educational tests, however, or when outside variables are correlated with test scores, we sometimes encoun-

[1] DuBois, P. H. Formulas and tables for rank correlations. *Psychol. Rec.*, 1939, **3**, 46–56.

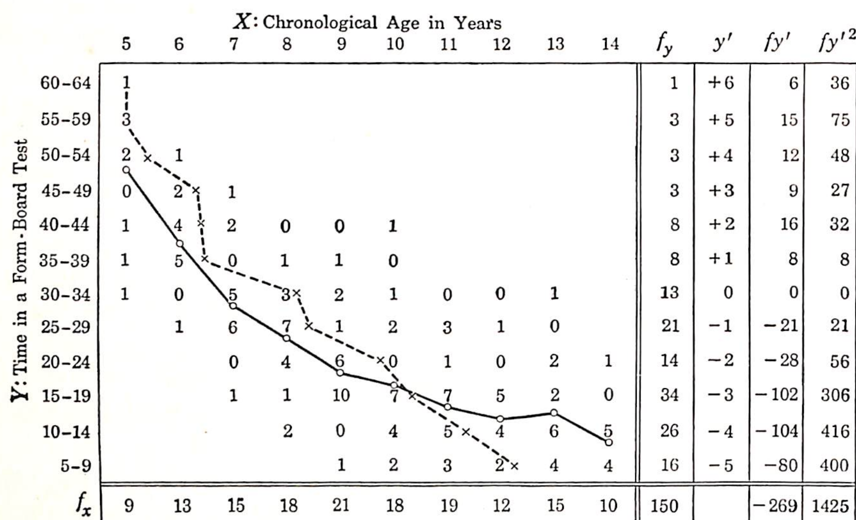
ter curved trends in the scatter diagram. The means of the columns do not progressively increase as we go up the X scale. They may increase slowly at first then rapidly later, or they may increase to a maximum in the center and then decrease, or other systematic divergencies from linearity may be apparent.

Nonlinear Regressions.—A common instance of nonlinear relationship is found when we correlate performance scores with chronological age. Typically, goodness of performance, as measured, increases most rapidly from ages five to ten and thereafter shows a slackening in upward trend through the teens. If we follow the progression still further, we find typically a maximal performance somewhere in the twenties, with slow decline to the forties and an increasing rate of decline thereafter. If we included all ages from five to seventy-five in our correlation study and if we computed the usual Pearson r between age and scores, the r would probably prove to be near zero. On such a correlation diagram, the scattering of points would be considerably dispersed from any straight line that we might try to draw through the data, slanting upward or slanting downward. Inspection would show, nevertheless, a law of relationship between age and performance but a relationship that takes into account the waxing and waning of ability both within the span of ages studied.

We might break the chart in two and treat by themselves the years during which there is improvement and by themselves the years during which there is decline. We should be able to compute a positive correlation for the earlier span and a negative correlation for the later span by assuming straight-line trends. But these would be of doubtful significance and certainly would not do justice to the full strength of relationships, even within the two segments of life span. The reason is that the trends still deviate from straight lines. Curvature has been overlooked, and to that extent the index of correlation is perhaps markedly underestimated.

Two Regression Lines and Two Correlation Ratios.—The scatter diagram in Fig. 13.1 represents a sample of relationship between performance score in a form-board test and chronological age between five and fifteen years inclusive. Here the score is time required for completion, so a high number indicates a poor performance, and the trend is downward. But the relationship obviously drops most rapidly during the first 3 years and settles down to slight changes from year to year during the last 3 years. Two regression lines are drawn in the diagram to show more clearly the trends. The regression of test score on age is shown by the solid line that is drawn connecting the circlets, which are plotted at the means of the columns. The regression of age upon test score is shown by

the dotted line and the means of the rows, by the x's. Just as we find two regression lines (for an imperfect correlation) in Ch. 15, where linear regressions are involved, so here we find two regression curves, differing in shape as well as in slope. We have accordingly two correlation ratios, or eta coefficients, one for each of the regressions, and they will not necessarily be the same in value. This result differs from that in the case of



Y	5	6	7	8	9	10	11	12	13	14	f_y	y'	fy'	fy'^2
60-64	1										1	+6	6	36
55-59	3										3	+5	15	75
50-54	2	1									3	+4	12	48
45-49	0	2	1								3	+3	9	27
40-44	1	4	2	0	0	1					8	+2	16	32
35-39	1	5	0	1	1	0					8	+1	8	8
30-34	1	0	5	3	2	1	0	0	1		13	0	0	0
25-29		1	6	7	1	2	3	1	0		21	−1	−21	21
20-24			0	4	6	0	1	0	2	1	14	−2	−28	56
15-19			1	1	10	7	7	5	2	0	34	−3	−102	306
10-14				2	0	4	5	4	6	5	26	−4	−104	416
5-9					1	2	3	2	4	4	16	−5	−80	400
f_x	9	13	15	18	21	18	19	12	15	10	150		−269	1425

FIG. 13.1.—A scatter diagram for a correlation-ratio problem.

linear correlation, where $r_{yz} = r_{xy}$. The two correlations ratios are given by the formulas

$$\eta_{yx} = \frac{\sigma_{y'}}{\sigma_y} \qquad \text{(Correlation ratio for the regression of } Y \text{ on } X) \qquad (13.4a)$$

and

$$\eta_{xy} = \frac{\sigma_{x'}}{\sigma_x} \qquad \text{(Same, for regression of } X \text{ on } Y) \qquad (13.4b)$$

where $\sigma_{y'}$ = standard deviation of the values (Y') predicted from X, and
$\sigma_{x'}$ = standard deviation of the X values predicted from Y.
σ_y and σ_x = standard deviations of the two total distributions.
The manner in which $\sigma_{y'}$ and $\sigma_{x'}$ are determined will next be explained.

The Computation of a Correlation Ratio.—In a prediction problem of this sort, the best prediction of Y for any column is the mean of the Y's in that column. This prediction will have the smallest sum of squared deviations from the observed Y's in that column. So Y' for each column is the mean of that column. We therefore first compute the means of the

columns. These are listed in column (3) in Table 13.2. Now if there were no correlation, no law of relationship between Y and X, these Y' values would lie along the level of the mean of *all* the Y values, which in this problem is 23.0. No predictions could then be made on the basis of knowledge of X values. For every column with its X value (midpoint), the most probable corresponding Y would be 23.0 and our margin of error would be indicated by σ_y. It would be as large as if we had no knowledge of X for each individual (see Ch. 15 for a more complete discussion of this point).

The more the means of the columns deviate from the mean of all the Y's the more accurate our predictions are. We are therefore interested in how far the Y' values do deviate from 23.0 in this problem. Those discrepancies $(Y' - M_y)$ are given in column (4) of Table 13.2. As usual, we square the discrepancies or deviations and find their mean as an indicator of how great is their average. The squared discrepancies $(Y' - M_y)^2$ are given in column (5) of Table 13.2. But before

TABLE 13.2.—THE COMPUTATION OF A CORRELATION RATIO FOR THE REGRESSION OF TIME SCORE ON CHRONOLOGICAL AGE

(1) X CA	(2) n_c	(3) Y' Time	(4) $Y' - M_y$	(5) $(Y' - M_y)^2$	(6) $n_c(Y' - M_y)^2$	
14	10	11.0	−12.0	144.00	1,440.00	
13	15	14.0	− 9.0	81.00	1,215.00	
12	12	14.5	− 8.5	72.25	867.00	
11	19	16.0	− 7.0	49.00	931.00	
10	18	18.1	− 4.9	24.01	432.18	
9	21	20.8	− 2.2	4.84	101.64	
8	18	25.1	+ 2.1	4.41	79.38	
7	15	31.3	+ 8.3	68.89	1,033.35	
6	13	40.5	+17.5	306.25	3,981.25	
5	9	49.8	+26.8	718.24	6,464.16	
Sum....	150	....	......	......	16,544.96	$\Sigma n_c(Y' - M_y)^2$
					110.2997	$\sigma^2_{y'}$
					10.50	$\sigma_{y'}$

finding a mean of the squared discrepancies, we weight each one for a column by the number of cases in that column. The weighed, squared discrepancy for each column will be found in the last column of Table 13.2. Then they are summed, and we divide by N, which is 150 in this problem, to find $\sigma^2_{y'}$, which is 110.2997. The square root of this is 10.50, which is the σ of the discrepancies.

Remember that these are *not* the discrepancies of the observed points from the predicted Y values, for the larger these are the *lower* our correlation. We are here interested in the size of discrepancies between predicted Y values and the mean of all Y values, and the *larger* these are the *higher* our correlation. When the correlation is perfect, $\sigma_{y'}$ is as large as σ_y, for then the ratio $\sigma_{y'}/\sigma_y$ equals 1.00. When $\sigma_{y'} = 0$, the ratio equals zero. In this problem, $\sigma_y = 12.535$. The correlation ratio is therefore

$$\eta_{yx} = \frac{10.50}{12.535} = .838$$

The steps in computing a correlation ratio may be summarized as follows. Remember that for finding η_{xy}, we are dealing with *rows* rather than columns, so the steps will be the same except for the substitution of the word *row* for the word *column* in what follows and the substitution of X for Y.

Step 1. Determine the mean of all the Y values and also their standard deviation.
Step 2. Determine the means of the columns (Y').
Step 3. Determine the discrepancies between Y' and M_y.
Step 4. Square the discrepancies.
Step 5. Multiply each squared discrepancy by the number of the cases in the column (n_c).
Step 6. Sum the weighted, squared discrepancies, and divide by N. This gives $\sigma^2_{y'}$. From this, find $\sigma_{y'}$.
Step 7. Solve the ratio $\sigma_{y'}/\sigma_y$, which is η_{yx}.

The Standard Error of a Correlation Ratio.—The reliability of a correlation ratio, like the reliability of r, is given by its standard error, and this is derived by a similar formula

$$\sigma_\eta = \frac{1 - \eta^2}{\sqrt{N - 1}} \qquad \text{(Standard error of a correlation ratio)} \quad (13.5)$$

The standard error of the eta coefficient that we have just obtained is .025. The amount of correlation is therefore rather close to the population correlation.

The Standard Error of Estimate in a Nonlinear Regression.—The standard error of estimate here can be computed as from a Pearson r (see formulas 15.16*a* and 15.16*b*), but it can also be obtained from the knowledge that

$$\sigma^2_{yx} + \sigma^2_{y'} = \sigma^2_y$$

That is, the total variance in the Y distribution is made up of two components, the variance predictable from X (this is $\sigma^2_{y'}$) and the variance not predictable from X (which is σ^2_{yx}). Transposing, we have

$$\sigma^2_{yx} = \sigma^2_y - \sigma^2_{y'}$$

In solving for an eta coefficient, we must know both the terms on the right of this equation. For our illustrative problem, they are 157.1262 and 110.2997, respectively. The difference is 46.8265, which is the non-predicted variance. The square root of this, which is 6.84, gives us σ_{yx}. The standard error of estimate tells us how much dispersion there is of the obtained values (Y values in this case) around the predicted values (Y' values in this case). The figure 6.84 tells us that two-thirds of the time scores in the Form Board test may be expected to be within 6.84 units of the predicted values, when the predicted values are the means of the columns of the scatter diagram.

The Relation of the Correlation Ratio to Analysis of Variance.—Those who have read Ch. 10 will find much that is familiar in the preceding paragraphs. Regarding the successive columns of data, which are really the result of a one-way classification on a quantitative variable, namely, chronological age, as sets, we have all the information we need to proceed with an analysis-of-variance solution (see Table 13.3). The sum 16,544.96

TABLE 13.3.—AN ANALYSIS OF VARIANCE BASED UPON STATISTICS DERIVED IN THE SOLUTION OF A CORRELATION RATIO

Component	Degrees of freedom	Sums of squares	Variances
Between sets	9	16,544.96	1,838.33
Within sets	140	7,023.97	50.17
Total	149	23,568.93	

$$F = \frac{1838.33}{50.17} = 36.6$$

will be recognized as the sum of squares between sets, since it is based upon the squared deviations of set means from the composite mean. The sum 7,023.97 will be recognized as the sum of squares within sets. This sum is found most conveniently here from what we already know. It is given by the product $N\sigma^2_{yx}$, which in this problem is $150 \times 46.8265 = 7023.97$. The sum of the two sums of squares makes up the total sum of squares for the composite sample in variable Y. All we need next are the degrees of freedom. For the between variance there are 9 (the number of sets minus

1). For the within variance there are 140 (N minus the number of sets). The two estimates of the population variance are given in Table 13.3, also the F ratio, which is 36.6. Reference to Table F (Appendix B) shows that it is well above the F required for significance at the 1 per cent level of confidence, which is about 2.5.

The relationship pointed out here is more of academic interest than of practical interest, for we already know that the eta coefficient was so high that there was little doubt of a law of relationship existing between chronological age and test score. Furthermore, the eta coefficient tells us a fact, namely, concerning the *degree* of relationship, which an F ratio does not convey. When the eta is near the lower margin of significance and a more rigorous test of significance is required, when a decision is to be made as to whether or not there is *any* genuine relationship at all, then the F test has its advantages.

A Test of Linearity of Regression.—Often the curvature in regression is so slight that we do not know but that it is merely a chance deviation from linearity. We therefore want some statistical test to show whether or not the curvature is probably real. Several tests of nonlinearity have been proposed. Probably the most dependable one is that suggested by Fisher. This method depends upon the already familiar chi square (see Ch. 11). For the solution of chi square here, we need to know the Pearson r for the same data for which an eta coefficient has been computed. The formula for chi square is

$$\chi^2 = (N - k)\left(\frac{\eta^2 - r^2}{1 - \eta^2}\right) \qquad \text{(Chi-square test of linearity)} \qquad (13.6)$$

where k = number of columns (or rows). For the problem in recent paragraphs, the Pearson r was found to be .763. By formula (13.6), we have

$$\begin{aligned}\chi^2 &= (150 - 10)\left(\frac{.702244 - .582169}{1 - .702244}\right)\\ &= (140)\left(\frac{.120075}{.297756}\right)\\ &= 56.5\end{aligned}$$

With a chi square of this size and $k - 2$ degrees of freedom, the divergence between η_{yx} and r_{yx} is so great as to leave little doubt about nonlinearity.

The hypothesis tested here is that the regression of Y and X is linear. In more exact terms, the hypothesis requires that the means of the columns all fall exactly along a straight line whose slope is described by the Pearson r. Now if the actual form of regression were linear, sampling errors would

cause the means of columns to deviate slightly from the best-fitting straight line. The sampling distribution is of these deviations of the actual means of the columns, the $M_{y'}$ values from the regression line. These deviations are ordinarily sufficient to make the eta coefficient larger than the Pearson r computed from the scatter diagram. The question is whether the deviations are large enough to suggest that there is something over and above these chance deviations involved. That is what the chi-square test here is supposed to tell us. The chi-square test should be applied to this particular use only when N exceeds k considerably.

An Evaluation of the Correlation Ratio.—The chief advantage and use of the eta coefficient has been indicated and illustrated—to determine the closeness of relationship between two variables when the regression is clearly nonlinear. Although very few nonlinear regressions have been found in the correlation of measures of ability, it is likely that there are many more such relationships in psychology and education than has been realized. This is true if we broaden our conception of the correlation problem considerably by saying that an *index* of correlation (index is a more inclusive term than coefficient) is a measure of the goodness of fit of obtained data to a regression line, whether it be straight or curved. The Pearson r indicates the goodness of fit of observed points to a straight line. Other indices, including eta, show the goodness of fit of data to other functions.

Correlation Coefficients as Indices of Goodness of Fit.—This broadening of the concept of correlation would bring into consideration curves of learning and retention and many others. The eta coefficient assumes no particular type of functional relationship between Y and X. The type of relationship is defined by the actual, unsmoothed trend of the means of the columns (or rows). In this fact is both strength and weakness. Allowing the curvature of the regression to be as complex as the ups and downs in obtained class means make it, we find in eta the maximum size of correlation index for any set of data. We might assume some kind of mathematical function for the data represented in Fig. 13.1—a hyperbola or parabola, a logarithmic function or some other. The goodness of fit, as indicated by a correlation index, would probably not be so high for any of these functions as the eta coefficient indicates. Because the eta coefficient does allow the regression curve to follow the means of the columns, a certain amount of error or purely sampling variance undoubtedly gets into the deviations of column means from the general mean of the Y's and hence the eta is a somewhat inflated figure. When the actual regression is linear, the difference between eta and r computed for the same data tells us about how much inflation has occurred. When the regression is nonlinear,

we have less ready evidence as to how much inflation there is. We should therefore discount any eta a little, particularly if the means of sets do not follow a smooth trend rather well. The smaller the sample, the more irregular the trend of the set means is likely to be, and therefore the greater the proportion of inflation in eta.

Examples of Nonlinear Regressions.—In addition to the functional relationships involved in learning and other phenomena, it is likely that when more is known about human traits that are not abilities—temperament, interests, and attitudes, and the like—and their interrelationships, we will find many more examples of nonlinear regression. In the validation of test scores against vocational or other criteria of adjustment, more and more of such examples are coming to light. It has been known for some time that high "intelligence" may be just as bad prognostically as low "intelligence" in connection with proficiency in routine and repetitive job assignments. This result will probably be found more general than has been supposed. The reason it has not been more widely recognized before is that relatively short ranges of ability have been related to proficiency criteria. If the total range, from lowest to the very highest, is studied in relation to proficiency indices on various kinds of jobs (except those requiring highest abilities) we may find the optimal ability to be somewhat short of the top in most cases. This definitely means nonlinear regressions. A number of instances have been called to the writer's attention in which scores on temperament tests bore a relation to rated proficiency in such a way that the optimal position on the trait score was barely above average. The application of the Pearson r method sometimes shows a near zero correlation in such instances whereas an eta coefficient might be as high as .30 or even .50. The straight line, in other words, was a very poor fit to the regression of the data. This should stress the importance of plotting scatter diagrams more frequently than is ordinarily done, otherwise important nonlinear regressions may be overlooked. No doubt many a zero Pearson r reported in the literature conceals a significant nonlinear relationship.

The Algebraic Sign of Eta.—Some writers regard it as a weakness of eta that its algebraic sign is always positive. The algebraic sign of r is meaningful in that it shows whether the general trend is upward or downward. In defense of eta it may be said that it tells us the thing we are most interested in, the goodness of fit or closeness of relationship between two things. If the over-all trend is either upward or downward we can readily perceive that by inspection of the scatter plot and we can attach whatever sign is appropriate if we wish to do so. Some curved regressions, *e.g.*, U-shaped or an inverted U-shaped type, may yield a significant eta without

any general trend away from the horizontal. In this case no sign is meaningful for eta.

Dependence of Eta upon the Number of Categories.—A more serious weakness of eta is that its size depends upon the number of columns (or rows). The minimum number of classes that would show any curvature at all is three, but three might give a much-smoothed and distorted view of the real relationship. With too small a number of classes, therefore, we run the chance of obtaining an estimate of correlation that is too small. On the other hand, as we increase the number of classes, we make the means of the classes less stable, and, as they fluctuate more, chance errors become more important in inflating eta. The limiting case would be classes so small that there was only one observation per class (assuming no duplicate measures on X), in which case the variance in the columns would be just as great as the over-all variance in Y and eta would equal 1.00. Methods for correcting eta for number of classes have been proposed, but none can be recommended. The best rule would be to keep the classes large enough so that means of classes are fairly stable and fall rather smoothly into line in the scatter plot and yet to have enough classes to bring out clearly enough the shape of the regression. The size of sample has some bearing on this. The larger the sample, the larger the number of classes that can be tolerated. Very small samples would be unsuitable for the computation of eta at all. With large samples (100 and above) it is suggested that the number of classes range between six and twelve.

The Use of Mathematical Functions.—Better than the correlation-ratio approach, in research studies, would be an effort to establish the form of a regression as some mathematical function and then test the goodness of fit of data to that function by methods which we cannot go into here. There are other texts that specifically treat this topic in some detail.[1]

The Biserial Coefficient of Correlation

The biserial r is especially designed for the situation in which both of the variables correlated are really continuously measurable but one of the two is for some reason reduced to two categories. This reduction to two categories may be a consequence of the only way in which the data can be obtained, as, for example, when one variable is whether or not a student passes or fails to pass a certain criterion of success. We can well assume a continuum along which individuals differ with respect to the ability required to pass this criterion. Those having a degree of ability

[1] Deming, W. E. Statistical adjustment of data. New York: Wiley, 1946; Lewis, D. Quantitative methods in psychology. Iowa City: The author, 1949.

above a certain crucial point do pass it, and those having a degree of ability below that crucial point fail to pass.

Let us assume that the criterion is graduation from pilot training. Although not all graduates are equal in achievement nor are all eliminees, all we know is whether each person belongs to one category or the other. It is as if our grouping were so coarse in this variable as to be confined to two class intervals rather than a dozen or so. If we are prepared to justify normality of distribution in this dichotomous variable, we have a formula by which a coefficient of correlation can be computed.

Computation of a Biserial r.—The principle upon which the formula for a biserial r is based is that with zero correlation there would be no

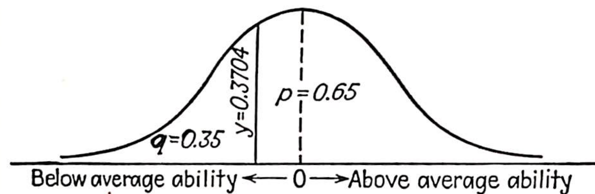


FIG. 13.2.—A normal distribution of the cases along the scale of ability to pass the course of training. The area to the right of the ordinate shown represents the 65 per cent who graduated and the area to the left represents the 35 per cent who failed to graduate.

difference between means, and the larger the difference between means the larger the correlation. The general formula for biserial r is

$$r_b = \frac{M_p - M_q}{\sigma_t} \times \frac{pq}{y} \quad \text{(Biserial coefficient of correlation)} \quad (13.7)$$

where M_p = mean of X values for the higher group in the dichotomous variable, the one having more of the ability in which the sample is divided into two subgroups.

M_q = mean of X values for the lower group.

p = proportion of the cases in the higher group.

q = proportion of the cases in the lower group.

y = ordinate of the normal distribution curve with surface equal to 1.00, at the point of division between segments containing p and q proportions of the cases (see Fig. 13.2).

σ_t = standard deviation of the total sample in the continuously measured variable, X.

Table 13.4 presents typical data for computing a biserial correlation. The passing group were distributed as shown; also, the failing group. The proportions passing and failing are .65 and .35, respectively.[1] The y

[1] It is good practice to compute p and q each to three significant digits.

TABLE 13.4.—DISTRIBUTIONS OF SCORES FOR TWO GROUPS OF STUDENTS—THOSE PASSING AND THOSE FAILING—ALSO A COMBINED DISTRIBUTION

	Scores											
	40–49	50–59	60–69	70–79	80–89	90–99	100–109	110–119	120–129	130–139	n	n/N
Passing students	...	1	3	10	27	30	26	21	7	5	130	$.65 = p$
Failing students	2	6	4	11	21	16	7	3	..	..	70	$.35 = q$
Total	2	7	7	21	48	46	33	24	7	5	200	1.00

ordinate (from Table C) is .3704. The distribution of the total group is assumed to be as indicated in Fig. 13.2. The computation of the biserial r proceeds as follows:

$$r_b = \frac{98.27 - 83.64}{17.68} \times \frac{(.65)(.35)}{.3704}$$
$$= \frac{3.328325}{6.548672}$$
$$= .508$$

Table G (Appendix B) is designed, in part, to supply several of the constants needed in the computation of a biserial r either by formula (13.7) or by formula (13.9), and the computation of its standard error. For given values of p, Table G supplies the corresponding values of pq/y, p/y, and $\sqrt{pq}/y$.

The Standard Error of r_b.—The standard error of a biserial r is estimated by the formula

$$\sigma_{r_b} = \frac{\frac{\sqrt{pq}}{y} - r^2_b}{\sqrt{N}} \quad \text{(Standard error of a biserial } r\text{)} \quad (13.8)$$

where the symbols have already been defined above.

In this problem

$$\sigma_{r_b} = \frac{\frac{.4770}{.3704} - .258064}{\sqrt{200}}$$
$$= \frac{1.0297}{14.142}$$
$$= .073$$

This standard error may be interpreted as usual, and we find that the obtained r_b is so large as undoubtedly not to be arising from an uncorrelated population.

Alternative Formula for Biserial *r*.—In many situations, a more convenient formula for the biserial *r* is[1]

$$r_b = \frac{M_p - M_t}{\sigma_t} \times \frac{p}{y} \qquad \text{(Alternative formula for a biserial } r\text{)} \qquad (13.9)$$

where the only new symbol is M_t, the mean of the total sample. The greater convenience of this formula over the other is that formula (13.9) gives us one less distribution to deal with. A good type of work sheet for solution by this formula is shown in Table 13.5. It is convenient to

TABLE 13.5.—SOLUTION OF MEANS AND STANDARD DEVIATION NECESSARY FOR THE COMPUTATION OF A BISERIAL *r*

Scores	x'	f_p	f_px'	f_t	f_tx'	$f_tx'^2$
130–139	+4	5	+20	5	+20	80
120–129	+3	7	+21	7	+21	63
110–119	+2	21	+42	24	+48	96
100–109	+1	26	+26	33	+33	33
90– 99	0	30	0	46	0	0
80– 89	−1	27	−27	48	−48	48
70– 79	−2	10	−20	21	−42	84
60– 69	−3	3	− 9	7	−21	63
50– 59	−4	1	− 4	7	−28	112
40– 49	−5			2	−10	50
Sums.	. . .	130	+49	200	−27	629

$c'_p = +.377 \qquad c'_t = -.135 \qquad \sigma_t = 10\sqrt{629/200 - .135^2}$

$c_p = +3.77 \qquad c_t = -1.35 \qquad = 10\sqrt{3.1268}$

$M_p = 98.27 \qquad M_t = 93.15 \qquad = 17.68$

use the same guessed mean for both the component distribution and for the total distribution. By this procedure, the biserial *r* and its σ_r come out the same as we have already seen.

An Evaluation of the Biserial *r*.—Since the biserial coefficient of correlation is a product-moment *r* and is designed to be a good estimate of the Pearson *r*, the same requirements as for the latter must be satisfied—linear regression and homoscedasticity—plus the unique requirement that the distribution of the values on the dichotomous variable, when continu-

[1] Dunlap, J. W. Note on computation of biserial correlations in item evaluation. *Psychom.*, 1936, **1**, 51–60.

ously measured, shall be normal. This requirement of normality applies to the form of population distribution. Even if the sample distribution is not normal, the population distribution may still be normal.

The use of the quantities p, q, and y in formulas (13.7) and (13.9) directly implies the normal distribution of the dichotomized variable. Departures from normality, if marked, will often lead to very erroneous estimates of correlation. With bimodal distributions, for example, it is possible that r will prove to exceed 1.0. Bimodal and other nonnormal distributions are most likely to occur in heterogeneous samples—for example, in variables on which there is a significant sex difference and both sexes are included in a sample.

When to Dichotomize Distributions.—The biserial r is very useful, in fact it is sometimes essential, and when properly used is a very good substitute for the Pearson r. There are instances in which the Y variable has been continuously measured but there are irregularities that preclude computing a good estimate of the Pearson r. In such cases the biserial r may be brought into service. One example of this would be a truncated distribution; another would be when there are very few categories for the Y variable and it is doubtful whether they are equidistant on a metric scale; another would be in the case of a badly skewed distribution in Y values owing to a defective measuring instrument. Before computing r_b, we would, of course, need to dichotomize each Y distribution. In this we would have some choice, and it would be well to make the division point as near the median as possible. The reason for this will be made clear in the next paragraph. In all of these peculiar instances, however, we are not relieved of the responsibility for defending the assumption of the normal distribution of Y. It may seem contradictory to suggest that when the Y distribution is skewed we resort to the biserial r, but note that it is the *sample* distribution that is skewed and it is the *population* distribution that must be assumed to be normal.

Biserial r Is Less Reliable than the Pearson r.—Whenever there is a real choice of computing a Pearson r versus a biserial r, however, one should favor the former, unless the sample is very large and unless computation time is an important factor. The standard error for a biserial r is quite a bit larger than that for a Pearson r derived from the same sample. If we compare the two formulas for the standard errors, formulas (9.24) and (13.8), we find that the only real difference is in the numerators. One reads $1 - r^2$ and the other reads $\sqrt{pq}/y - r^2_b$. If we examine the $\sqrt{pq}/y$ values in Table G, we find that even when this value is smallest (and that is when $p = q = .5$), it is about 25 per cent larger than 1. When $r_b = .00$, the standard error of r_b is therefore at least 25 per cent larger than that

for r for the same size of sample. As p approaches 1.0 or 0.0, the ratio $(\sqrt{pq}/y)$ becomes larger, until when $p = .94$, it is as large as 2. This is why in the preceding paragraph it was recommended that dichotomies have the division point as near the median as possible. It also suggests that we need larger samples for the same dependability of r_b than for r and that we should hesitate to compute r_b for very one-sided divisions of cases unless the sample is extremely large. This is reasonable from another point of view. Remember that prominent in the formula for r_b is the difference between means. This difference is not very stable unless each mean comes from a sample of sufficient size. Even if the sample totaled 1,000 cases, if only 1 per cent of the cases were in one of the two categories, its mean would be based upon only 10 cases. This is not favorable to reliable estimates based upon such a mean.

Other Serial Correlations.—Formulas have recently been developed by Jaspen for the correlation of a continuous variable with another variable that has been artificially classified in three, four, or five categories.[1] Owing to the rareness of the need for such formulas space will not be taken to present them here. If one has more than two categories, he can always combine certain ones to make two and then compute r_b, provided, of course, the necessary assumptions are satisfied.

Point-biserial Correlation

When one of the two variables in a correlation problem is a genuine dichotomy, or when it is doubtful that the dichotomous one stems from a normal distribution, the appropriate type of coefficient to use is the point-biserial r. Examples of genuine dichotomies are male versus female, being a farmer versus not being a farmer, owning a home versus not owning one, living versus dying, living in Boston versus not living in Boston, and so on. Bimodal or other peculiar distributions, although not representing entirely discrete categories, are sufficiently discontinuous to call for the point-biserial rather than the biserial r. Examples of this type are color blindness versus normal color vision; being alcoholic versus nonalcoholic; and criminal versus noncriminal.

There are other variables, though not fundamentally dichotomous and they may even be normally distributed, which we have to treat as if they were genuine dichotomies in practical operations. An outstanding example of this is a test item that is scored as either right or wrong. No doubt those who answer the item correctly are not all equally capable in the ability or abilities measured by the item. A total test score would provide continuous gradations in ability levels. In testing practice, how-

[1] Jaspen, N. Serial correlation. *Psychom.*, 1946, **11**, 23–30.

ever, the kind of item described is limited to separating individuals into two groups, and only gross predictions can be made from responses to it. Such a variable is a good example to explain the basic nature of the point biserial r. If we gave a "score" of +1 to each person with a correct answer and a "score" of zero to each person with a wrong answer, in the item variable we would have only two class intervals and we treat them as if they were genuine intervals. A product-moment r could be computed with Pearson's basic formula. The result would be a point-biserial r. A special formula is provided, however, which does not resemble the basic Pearson formula. It reads,

$$r_{pbi} = \frac{M_p - M_q}{\sigma_t}\sqrt{pq} \quad \text{(The point-biserial coefficient of correlation)} \qquad (13.10)$$

where the symbols are defined just as they were in the formula for the ordinary biserial r (formula 13.7).[1] The only difference between this formula and the one for the ordinary biserial r is that the numerator contains $\sqrt{pq}$ rather than pq, and the constant y is missing from the denominator. For the same set of data, then, the ordinary biserial r would be $\sqrt{pq}/y$ times as large as r_{pbi}. In this ratio lies a feature of r_{pbi} to which we will return soon.

Let us apply formula (13.10) to some data on the relation of body weight to sex membership. In a sample of 51 sixteen-year-old high-school students, of whom 24 were male and 27 were female, the mean weights in kilograms were 67.8 and 56.6, respectively. The proportion of males is accordingly 24/51 = .471 and q is .529. The standard deviation of the combined distributions was 13.2. This is a rather small sample on which to compute a point-biserial r, but it is favorably divided near the median and, at any rate, will do as a simple illustration. Solving with formula (13.10),

$$\begin{aligned} r_{pbi} &= \frac{67.8 - 56.6}{13.2}\sqrt{(.471)(.529)} \\ &= \frac{11.2}{13.2} \times .499 \\ &= .42 \end{aligned}$$

The correlation between sex and body weight for sixteen-year-old high-school students is estimated to be .42.

To the knowledge of the author no standard-error formula has been developed for r_{pbi}. It is suggested that a test of the hypothesis of zero correlation can be made by means of a t test of the difference $M_p - M_q$.

[1] For a derivation of this formula, also formula (13.11), see Appendix A.

The decision about this hypothesis should be the same as that for the hypothesis of a zero difference.

Alternative Methods of Computation for r_{pbi}.—As for the ordinary biserial r, there is an alternative formula for computing r_{pbi}, which may be more convenient in many situations. It reads

$$r_{pbi} = \frac{M_p - M_t}{\sigma_t}\sqrt{\frac{p}{q}} \quad \text{(Alternative formula for the point-biserial correlation coefficient)} \qquad (13.11)$$

Formulas for r_{pbi} making unnecessary the computation of p and q are

$$r_{pli} = \frac{(M_p - M_q)\sqrt{N_pN_q}}{N\sigma_t} \quad \text{(Alternative formulas for the point-biserial } r\text{)} \qquad (13.12)$$

and

$$r_{pbi} = \frac{(M_p - M_t)}{\sigma_t}\sqrt{\frac{N_p}{N_q}} \qquad (13.13)$$

where N_p and N_q are the frequencies in the two categories.

An Evaluation of the Point-biserial r.—Since the r_{pbi} coefficient is not restricted to normal distributions in the dichotomous variable, it is much more generally applicable than is r_b. Whenever there is doubt about computing r_b, the point-biserial r will serve. For this reason, it should probably be used more than it is. Being a product-moment r, it rests upon the assumption of linear regression, though when the dichotomy is genuine, the regression could be nothing other than linear. Although a product-moment r, in value it is rarely comparable numerically with a Pearson r, or even with an ordinary biserial r, even when computed from the same data. This is its greatest weakness as a descriptive statistic. Under special circumstances, to be described, it may be used as a basis for making an estimate of the Pearson r.

Relation of r_{pbi} to r_b.—When properly applied, r_b gives coefficients that are generally good approximations to Pearson r's that could be computed from the same data had both variables been continuously measured. Consequently, all the usual interpretations that are made of r (see Ch. 15) can also be made of r_b.

If r_{pbi} were computed from data that actually justified the use of r_b, however, the coefficient computed would be markedly smaller than r_b obtained from the same data. Even if the one variable is actually continuous but not normally distributed, in which case we might better utilize r_{pbi}, the latter would give an underestimate of the amount of correlation. As was pointed out before, r_b is $\sqrt{pq}/y$ times as large as r_{pbi} when they are computed from the same basic data. This ratio varies

from about 1.25 when $p = .50$ to about 3.73 when p (or q) equals .99 (see Table G). Fig. 13.3 shows graphically the ratio of r_{pbi} to r_b for various values of p. The ratio of r_{pbi} to r_b is, of course, the reciprocal of the ratio of r_b to r_{pbi}, in other words, it is $y/\sqrt{pq}$. The diagram is designed in this

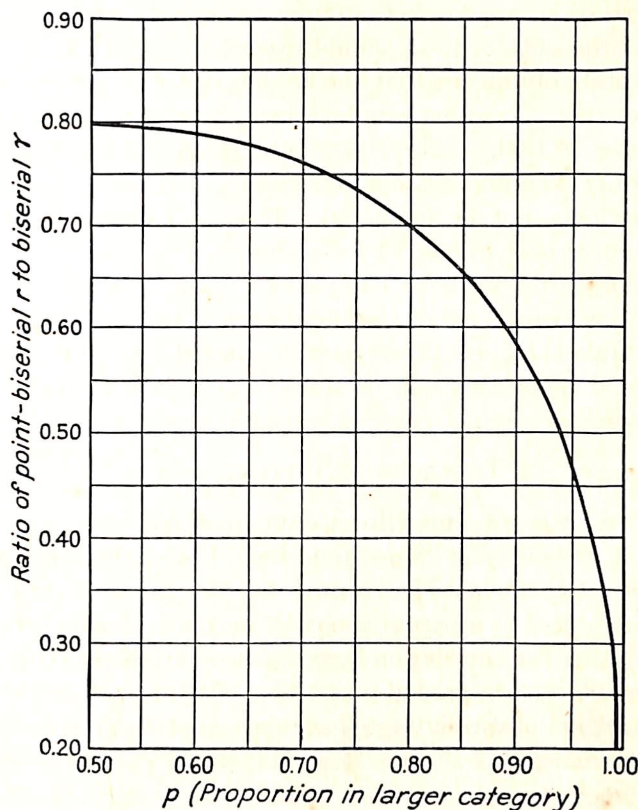


Fig. 13.3.—Ratio of the point-biserial r to the biserial r when the difference between means ($M_p - M_q$) and the standard deviation (σ_t) are constant and the proportion in the larger category (p) varies.

manner to show maximum values of r_{pbi} that would arise from continuous, normal distributions. In terms of formulas,

$$r_b = r_{pbi} \frac{\sqrt{pq}}{y} \tag{13.14a}$$

(Conversion of one biserial r into the other when normality of distribution exists)

$$r_{pbi} = r_b \frac{y}{\sqrt{pq}} \tag{13.14b}$$

It is recommended that when the dichotomous variable is normally distributed without much doubt, r_b be computed and so interpreted. If there is little doubt that the distribution is a genuine dichotomy, r_{pbi}

should be computed and so interpreted. For the doubtful situations, the r_{pbi} should be computed but interpreted in the light of Fig. 13.3. That is to say, if the distribution in question is continuous but not normal, and if r_{pbi} approaches the limit described by Fig. 13.3, we can say that the genuine correlation approaches 1.00 more closely than the obtained r_{pbi} does. If the obtained r_{pbi} should exceed the limit, for the size of p involved, it probably means that the assumption of a genuine dichotomy is the correct one. In other words, when there is a point distribution, r_{pbi} can approach 1.00. Many distributions are in the doubtful class; they are neither dichotomous nor continuous. At least, if they are continuous, they may not be unimodal. It is to help take care of these twilight instances that Figure 13.3 was designed.

If it develops after we have computed r_{pbi} that the situation justifies the use of r_b, we can convert the obtained r_{pbi} to the appropriate r_b by means of formula (13.14*a*). If we have computed r_b when it later develops that we should have used r_{pbi}, formula (13.14*b*) will provide the proper transformation.

Tetrachoric Correlation

A tetrachoric r is computed from data in which both X and Y have been reduced artificially to two categories. Under the appropriate conditions it gives a coefficient that is numerically equivalent to a Pearson r and may be regarded as an approximation to it. It is sometimes the only way of estimating the correlation between two variables because the data could not be obtained in graded quantities. It is sometimes a quick and convenient method of estimating r from data that are in the form of continuous measurements, but time is an important consideration and the sample is large.

Assumptions Underlying the Tetrachoric r.—The tetrachoric r requires that both X and Y be continuously variable, normally distributed, and linearly related. A problem in which the tetrachoric r may be computed is illustrated in Table 13.6, if we are willing to make the necessary assumptions. These data represent the numbers of students responding "Yes" and "No" to two questions in a personality questionnaire. Question I was, "Do you enjoy getting acquainted with most people?" and Question II was, "Do you prefer to work with others rather than alone?" Out of 930 replies to both questions, we have the numbers who responded similarly (cells a and d in Table 13.6) and the number who responded differently to the two questions (cells b and c). It is obvious that in the case of a perfect positive correlation, all the cases would fall in cells a and d. In a perfect negative correlation, they would fall in cells b and c.

TABLE 13.6.—FOURFOLD TABLE FROM WHICH A TETRACHORIC COEFFICIENT OF CORRELATION IS COMPUTED

Question II	Question I: Yes	No	Total	Proportion	Division ordinate	Point deviate
Yes	374 (a)	167 (b)	541	.582 (p)	.3905 (y)	.2070 (z)
No	186 (c)	203 (d)	389	.418 (q)		
Total	560	370	930	1.000		
Proportion	.602 (p')	.398 (q')	1.000			
Ordinate	.3858 (y')					
Deviate	.2585 (z')					

In a zero correlation, the frequencies would be proportionately distributed in the four cells.[1]

The assumption of continuity and normality of distribution can be defended as follows: It is unlikely that all who respond "Yes" to either question do so with equal degree of affirmation. It is similarly unlikely that those who respond "No" do so with equal degree of negation. It is most likely that either question represents a continuum of behavior extending from strong affirmation at the one extreme to strong negation at the other. Continuity is thus the probable state of affairs, not a real dichotomy. If a continuum is granted, the general law of unimodal distribution approaching normality in psychological traits may be cited in defense of the other requirement. By making the necessary assumptions, at any rate, many things can be done with such data that would otherwise be impossible. As in most statistical operations where true form of distribution is unknown, we can here remember that we have taken the chance of faulty assumptions and interpret results with the requisite reservation.

The Equation for the Tetrachoic r.—The complete equation for the tetrachoric r is indeed a long and complicated one, involving a series including many of powers of r. The first few terms included, it reads

[1] It will be noted that the categories for X are in an unusual order (positive or "good" end toward the left), which makes the regression "line" slope downward to the right for a positive correlation. For some reason, tradition has kept to this arrangement. Other 2×2 tables reverse this order, in keeping with the usual scatter diagram. Then the letters a and b, also c and d, are reversed. Letters a and d always stand for like-signed cases in this volume.

$$r_t + r^2_t \frac{zz'}{2} + r^3_t \frac{(z^2 - 1)(z'^2 - 1)}{6} + \cdots = \frac{ad - bc}{yy'N^2} \qquad (13.15)$$

The symbols will be explained with reference to Table 13.6. The letters a, b, c, and d refer to the frequencies in the four cells of the fourfold table. r_t is given the subscript to indicate that it is a tetrachoric r. Numerically, it is equivalent to a Pearson r.

In Table 13.6, it will be noted that the distribution of responses to Question I is given in terms of proportions p' and q'. The distribution of all responses to Question II is similarly given in terms of p and q. These proportions are required for finding the values for the y's and z's in formula (13.15). The symbols z and z' stand for the standard measurements on the base line of the normal distribution curve at the points of division of cases in the two distributions.

From Table C, we find that z is .2070 and z' is .2585. The symbols y and y' stand for the ordinates in the normal curve at the points of division. From Table C, we find that they are .3905 and .3858, respectively. N is, of course, 930. We now have all the values except r_t, for which we must solve the equation.

An Approximate Solution for a Tetrachoric r.—Let us ignore all terms involving higher powers of r_t than the second. We can then reduce formula (13.15) to a quadratic equation that is readily soluble but that will give only an approximation to r_t. The terms ignored are rather small, however, and so can be disregarded. With substitutions, the equation becomes

$$r_t + \frac{(.2070)(.2585)}{2} r^2 = \frac{(374)(203) - (167)(186)}{(.3905)(.3858)(930^2)}$$

which reduces to

$$r_t + .026755r^2_t = .344279$$

It is well in this solution to carry at least six decimal places in order to assure a sufficient number of significant digits later.

We have now arrived at a quadratic equation, which, with rearrangement of terms, becomes

$$.026755r^2_t + r_t - .344279 = 0$$

And this is in a form to which we can readily apply a standard algebraic solution. If the standard quadratic equation is written

$$ar^2_t + br_t + c = 0$$

we can solve for r_t by using the following formula:

$$r_t = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a} \tag{13.16}$$

In our equation for r_t above, a is .026755, b is 1.0, and c is $-.344279$. Substituting these in formula (13.16) we have

$$r_t = \frac{-1.0 \pm \sqrt{1 - 4(.026755)(-.344279)}}{2(.026755)}$$
$$= \frac{-1.0 \pm 1.01825}{.05351}$$

Here we have a choice between the positive or the negative square root. If we choose the negative one, the numerator becomes -2.01825, which would give us an r_t equal to about -40. This is obviously absurd, since no law-abiding r can go below -1. Taking the positive root, we have

$$r_t = \frac{.01825}{.05351}$$
$$= .341, \text{ or } +.34$$

We can now say that, if our assumptions about the two distributions are granted, the correlation between an expressed enjoyment of getting acquainted with people and an expressed preference for working with others is $+.34$. For greater refinement in the solution, we could now treat .341 as a trial value for r_t in equation (13.15) and see how much discrepancy is involved when the term having r^3 in it is included in the calculations. We could make any change in r_t that seemed necessary for a better satisfaction of the equation and by successive trial-and-error maneuvers arrive at a more exact choice of r_t. Probably most data are not of sufficient number or precision to justify the extra labor involved in this. The discrepancy involved when all powers higher than two are ignored in equation (13.15) is probably much smaller than the standard error of r_t, unless r is fairly large.

The Standard Error of a Tetrachoric r.—The tetrachoric r is less reliable than the Pearson r, being as much as 50 per cent more variable. r_t is most reliable (1) when N is large, as is true of all statistics, (2) when r is large, as is true of other r's, but also (3), when the divisions into two categories are close to the medians. The complete formula for estimating σ_{r_t} is entirely too long to be practical; so it will not be given here. But when $r_t = .0$, the formula is much simpler and reads

$$\sigma_{r_t} = \frac{\sqrt{pp'qq'}}{yy'\sqrt{N}} \quad \text{(Standard error of a zero tetrachoric } r\text{)} \tag{13.17}$$

where the symbols mean the same as in formula (13.15) or in Table 13.6.[1] This is the most useful of the formulas for σ_{r_t}, at any rate, because it permits testing the null hypothesis. If the correlation in the population is zero, samples of the size we obtained would yield r's with a standard error as given by this equation. For the 930 cases in our problem,

$$\sigma_{r_t} = \frac{\sqrt{(.582)(.602)(.418)(.398)}}{(.3905)(.3858)\sqrt{930}}$$
$$= .053$$

Since the obtained r_t is .34, being more than 2.6 times this standard error, we can be quite positive that the two qualities represented by the two questions are really correlated in the population. Had this been a Pearson r, σ_r (for $r = .0$) would have been .033. This gives a rough idea of the relative degree of sampling fluctuation of the two kinds of r and also bears out the statement made earlier that the tetrachoric r is about 50 per cent more variable. This fact should impress one with the importance of using a larger sample when r_t is to be the index of correlation. Roughly, to attain the same degree of reliability in a tetrachoric r, one needs more than twice the number of cases as in the use of the Pearson r. For very dependable results, it is recommended that N be at least 200 and preferably 300 when r_t is to be computed. In smaller samples than these, even less than $N = 100$, a tetrachoric r can be used to test the null hypothesis, but it cannot be depended upon to give very accurate estimates of the size of correlation unless r is very large.

Other Procedures for Estimating r_t.—A tetrachoric r can be estimated well enough for practical purposes by a number of short-cut procedures, two of which will now be described.

The Cosine-pi Formula.—One approximation formula for r_t is known as the cosine-pi formula. In mathematical form,

$$r_t = \cos\left(\pi \frac{\sqrt{bc}}{\sqrt{ad} + \sqrt{bc}}\right)$$

Since for computing purposes here π can be taken as 180 degrees, the practical form of the equation is

$$r_t = \cos\left(\frac{180° \sqrt{bc}}{\sqrt{ad} + \sqrt{bc}}\right) \quad (13.18)$$

(Cosine-pi approximation formula for estimating a tetrachoric r)

[1] For aids in estimating σ_{r_t}, see Guilford, J. P. and Lyons, T. C. On determining the reliability and significance of a tetrachoric coefficient of correlation, *Psychom.*, 1942, **7**, 243–249; also Hayes, S. P. Tables of the standard error of tetrachoric correlation coefficient, *Psychom.*, 1943, **8**, 193–203.

where a, b, c, and d, are the frequencies as defined in Table 13.6. It is well to remember that b and c represent the unlike-signed cases and a and d the like-signed cases. When numbers are substituted, the expression within the parentheses reduces to a single number which is an angle in terms of degrees of arc. The cosine of this angle is the estimate of r_t. The angle will vary between zero, when either b or c is zero, or both, to 180°, when either a or d is zero, or both. In the first case, when the angle is zero, the correlation is $+1.0$ and in the second case, when the angle is 180°, r_t is -1.0. When the product bc equals ad, the angle is 90°, the cosine of which is zero, and $r_t = .0$.

Let us apply the cosine-pi formula to the data of Table 13.6.

$$\begin{aligned} r_t &= \cos\left[\frac{180° \sqrt{(167)(186)}}{\sqrt{(374)(203)} + \sqrt{(167)(186)}}\right] \\ &= \cos\left[\frac{180°(176.3)}{275.5 + 176.3}\right] \\ &= \cos 70.24° \\ &= .343 \end{aligned}$$

The cosine of an angle of 70.24 degrees (as found by interpolating in Table J, Appendix B) is .343. This estimate of r_t for these data checks very closely with that reported earlier (.341).

In this method, if the angle should prove to be between 90° and 180° the correlation is negative. This can be anticipated by noting that the product bc is greater than ad. Angles over 90° are not listed in Table J. For angles between 90° and 180°, find the cosine of 180° minus the obtained angle.

Graphic Estimates of Tetrachoric r.—When a large number of tetrachoric r's must be computed, considerable saving of labor is provided by the Thurstone computing diagrams.[1] These are highly recommended since they yield two-place accuracy with little effort after the fourfold table is completely reduced to the status of proportions throughout, as in Table 13.7. From the computing diagrams, r_t for the data in Table 13.7 is estimated to be $+.79$. The correlation of the two questions of Table 13.6 is estimated as $+.34$, which checks with previous estimates. Another graphic procedure was recently published by Hayes.[2]

[1] Chesire, L., Saffir M., and Thurstone, L. L. Computing diagrams for the tetrachoric correlation coefficient. Chicago: University of Chicago Bookstore, 1938.

[2] Hayes, S. P. Diagrams for computing tetrachoric correlation coefficients from percentage differences, *Psychom.*, 1946, **11**, 163–172.

TABLE 13.7.—THE REDUCTION OF A SCATTER DIAGRAM TO A FOURFOLD TABLE PREPARATORY TO THE COMPUTATION OF A TETRACHORIC COEFFICIENT OF CORRELATION*

Mark in Schoolwork

IQ	F	D	C	B	A	Total
120 and above	..	...	12	32	40	84
110–119	..	4	23	66	23	116
100–109	1	10	67	77	15	170
90– 99	1	22	133	40	3	199
80– 89	8	71	125	21	2	227
70– 79	36	92	24	1	..	153
Below 70	27	36	4	...	..	67
Total	73	235	388	237	83	1,016

IQ	In terms of frequencies			In terms of proportions		
	C, or below	A or B	Total	C, or below	A or B	Total
90 or above	273	296	569	.269	.291	.560
Below 90	423	24	447	.416	.024	.440
Total	696	320	1016	.685	.315	1.000

* Adapted from Cobb, M. V. The limits set to educational achievement by limited intelligence. *J. educ. Psychol.*, **13**, 1922, p. 449. By permission of the publisher.

Reducing Distributions in Class Intervals to Fourfold Tables.—Data need not be obtained in two categories each way in order to apply the tetrachoric solution for r. Any scatter diagram, in fact, can be reduced to two groups each way by making arbitrary divisions. Such a division should be made as nearly as possible at or near the median in each distribution. Table 13.7 shows a scatter diagram in which reduction to a fourfold table would be highly desirable. A Pearson r computed with so few class intervals each way would be highly influenced by errors of grouping. The very large number of cases renders the reduction in reliability of r by computing r_t of small importance. The divisions suggested in Table 13.7 come between the *B*'s and *C*'s for distribution of school marks and at an *IQ* of 90 for intelligence rating. The revised correlation distribution is seen in Table 13.7.

Some Applications of r_t to Be Avoided.—Many of the limitations of the tetrachoric r have already been pointed out. There are others which should not go unnoticed. It is well to avoid estimating r_t when the split in either X or Y is very one-sided—for example, a 95-5, or even a 90-10, division of the cases. The standard error is much larger in such situations as these.

Especially to be avoided is an attempt to estimate r_t when there is a zero in only one cell. Table 13.8, *A* and *B*, illustrates two such examples.

TABLE 13.8.—ILLUSTRATIONS OF SOME UNUSUAL FOURFOLD CONTINGENCY TABLES IN WHICH COMPUTATION OF A TETRACHORIC *r* IS QUESTIONABLE

A

0	200	200
110	90	200
110	290	400

B

110	80	190
0	150	150
110	230	340

C

15	85	100
105	95	200
120	180	300

If r_t were computed for problem *A* it would equal -1.0 (the zero is in cell a); if computed for problem *B*, r_t would equal $+1.0$. This is in spite of the fact that about one-fourth of the cases belie the perfect correlations apparent by computation (90 cases out of 400 in *A* are out of line with the finding and 80 cases in *B*). These examples are perhaps somewhat rare, but zero frequencies are certainly not unheard of. Even scatters like that in *C* would probably give a false estimate of correlation. There is no zero, but there is an exceptionally small frequency (15) among much larger ones. In all three fourfold tables the distributions are such as to suggest nonlinear regressions if these broad categories were broken down into finer groupings. If the assumption of linearity is not satisfied, r_t may well give a biased estimate of correlation. Such distributions as those in Table 13.8 are not proof of nonlinear regression but they strongly suggest it. In general, a distribution in such a table should appear to be rather symmetrical around one diagonal axis or the other, depending upon whether the correlation is negative or positive. This holds true if the proportion p is somewhat near the proportion p' but if they differ too much, asymmetry cannot be taken necessarily to mean curved regression.

THE PHI COEFFICIENT

When the two distributions correlated are really dichotomous, when the two classes are separated by a real gap between them and previous correlational methods do not apply, we may resort to the phi coefficient.[1] This was designed for so-called *point distributions*, which implies that the two classes have two point values or merely represent some unmeasurable attribute. Such a case would be illustrated by eye color, sex, "living versus dead," and the like. The method can be applied, however, to data that are measurable on a continuous variable if we make certain allow-

[1] Also known as the Yule ϕ or sometimes as the Yule-Boas ϕ. See Yule, G. U. On the methods of measuring the association between two attributes. *J. Roy. Stat. Soc.*, 1912, **75**, 576–642.

ances for that fact. It is a close relative of chi square, which is applicable to a wide variety of situations.

The Computation of Phi.—To illustrate the use of phi (ϕ), we shall use again some data that were previously employed with chi square (see Table 11.3). They are repeated here as we need them, in proportion form, in Table 13.9.

TABLE 13.9.—A TABLE TO ILLUSTRATE THE CORRELATION OF ATTRIBUTES

	Normal	Feebleminded	Both
Married	.269 (α)	.204 (β)	.473 (p)
Unmarried	.231 (γ)	.296 (δ)	.527 (q)
Both	.500 (p')	.500 (q')	1.000

The formula for the phi coefficient is

$$\phi = \frac{\alpha\delta - \beta\gamma}{\sqrt{pqp'q'}} \qquad \text{(The phi correlation coefficient)} \qquad (13.19)$$

where the symbols correspond to the labeled cells in Table 13.9.[1] The solution of ϕ for this table is

$$\phi = \frac{(.269)(.296) - (.204)(.231)}{\sqrt{(.473)(.527)(.5)(.5)}}$$
$$= \frac{.0325}{.2496}$$
$$= .1302, \text{ or } .13$$

The Relation of Phi to Chi Square.—Phi is related to chi square from a 2 × 2 table by the very simple equation

$$\chi^2 = N\phi^2 \qquad \text{(Chi square as a function of phi)} \qquad (13.20)$$

and phi is derived from chi square by the equation

$$\phi = \sqrt{\frac{\chi^2}{N}} \qquad \text{(Phi as a function of chi square)} \qquad (13.21)$$

By formula (13.20)

$$\chi^2 = (412)(.016952)$$
$$= 6.98$$

This checks with the solution of chi square by other methods (see Ch. 11).

[1] For a derivation of formula (13.19), see Appendix A.

Since phi can be derived directly from chi square, when applied to a 2×2 table, any of the formulas for chi square given in Ch. 11 will apply to its computation. Formula (11.7), especially, which is very similar to formula (13.19) above, is probably more convenient. Applied directly to the computing of phi, it becomes

$$\phi = \frac{(ad - bc)}{\sqrt{(a + b)(a + c)(b + d)(c + d)}} \quad \text{(Phi computed from frequencies)} \qquad (13.22)$$

The Special Case of Phi When One Distribution Is Evenly Divided.—When one of the distributions, let us say the one for which we use p' and q' as total proportions, is evenly divided so that $p' = q' = .50$, the solution of ϕ is considerably simplified. The formula reads

$$\phi = \frac{\alpha - \beta}{\sqrt{pq}} \quad \text{(Phi from evenly divided proportions)} \qquad (13.23)$$

Applied to the data on marital status

$$\phi = \frac{.269 - .204}{\sqrt{(.473)(.527)}} = \frac{.065}{.4993} = .130$$

This particular case is useful in many an experimental situation where two separated groups are selected with equal numbers of cases. There is some question here, of course, as to how well the samples chosen represent the larger population from which they were obtained, and so interpretations should be stated with this knowledge in mind.

The Reliability and Significance of Phi.—The formula for the estimation of the standard error of phi involves such laborious computations that it is impractical for general use. It will not be given here. A test of the null hypothesis, fortunately, can be made through phi's relationship to chi square. If χ^2 is significant in a fourfold table, the corresponding ϕ is significant. The procedure, then, is to derive the corresponding χ^2 from the obtained ϕ by means of formula (13.20), then examine Table E to find whether for one degree of freedom the required standard of significance is met. In the marital problem, we find that a chi square of 6.98 is significant beyond the 1 per cent level, therefore the obtained phi of .13 is likewise significant.

An Evaluation of the Phi Coefficient.—Phi is actually a product-moment coefficient of correlation. Its formula is a variation of Pearson's funda-

mental equation, $r = \Sigma xy/N\sigma_x\sigma_y$. The similarity may be seen to some degree, at least, if we break the denominator of formula (13.19) into two components, $\sqrt{pq}$ and $\sqrt{p'q'}$. These are the standard deviations of the two point distributions, in Y and X. If we give numerical values of +1 and 0 to the two categories in X and in Y, and if we carry through the computation of a Pearson r in a scatter diagram of four cells, we arrive at a correlation coefficient equal to ϕ.

Limitations to the Size of Phi.—While ϕ can vary from −1.0 to +1.0, only under certain conditions can ϕ be as large as either of these extremes, even though a tetrachoric r, if computed for the same data, would yield an r_t equal to 1. This is probably its greatest weakness, but in certain practical situations (see Ch. 17) it is a realistic feature. The reason is that the reduction of frequencies to a 2 × 2 table places serious restrictions upon ϕ that do not affect r_t. The general principle is that ϕ can be as great as 1 only when the two variables are divided so that $p = p'$ or $p = q'$ (and, of course, $q = q'$ or $q = p'$). To illustrate these restrictions, we may take a few special cases in which $p = .5$ but p' is allowed to vary. Such instances are pictured in Table 13.10. With an even division of the

TABLE 13.10.—SOME FOURFOLD CONTINGENCY TABLES ILLUSTRATING THE DEPENDENCE OF THE SIZE OF A PHI COEFFICIENT UPON THE MARGINAL TOTALS

		+	−	
	+	50	0	50
Y	−	0	50	50
		50	50	100

X

$\phi = +1.0$

A

		+	−	
	+	0	50	50
Y	−	50	0	50
		50	50	100

X

$\phi = -1.0$

B

		+	−	
	+	50	0	50
Y	−	25	25	50
		75	25	100

X

$\phi = .58$

C

		+	−	
	+	50	0	50
Y	−	40	10	50
		90	10	100

X

$\phi = .33$

D

		+	−	
	+	45	5	50
Y	−	30	20	50
		75	25	100

X

$\phi = .35$

E

cases in the two categories in Y, only with an even division also in X is it possible to have a perfect correlation, as shown in contingency tables A and B. With a division of 75-25 in variable X, the maximum ϕ would be .58 (contingency table C) and with a 90-10 division, the maximum ϕ would be .33. In contingency table E the division in X is again 75-25 but there is departure from maximal relationship. The obtained ϕ of .35 may be interpreted for size in the light of the maximal ϕ possible with the particular combination of marginal totals, if we are interested in the underlying strength of relationship between X and Y. If we are interested in making predictions from categories to other categories, however, the obtained ϕ is a more realistic figure. The problems of prediction come in the chapters to follow.

Determination of a Maximal Phi Coefficient.—Because of the increasing importance of the phi coefficient, particularly in connection with test-item intercorrelations, it is desirable for the purposes of orientation to have some conception of the drastic limitations to the size of phi. In general, the maximal ϕ for any combination of p and p' can be calculated by means of the formula

$$\phi_{max} = \sqrt{\left(\frac{p_j}{q_j}\right)\left(\frac{q_i}{p_i}\right)} \text{ (where } p_i \geqq p_j \geqq .5) \quad \text{(Maximal value for } \phi \text{ with different combinations of } p_i \text{ and } p_j) \quad (13.24)$$

where p_i = largest marginal proportion in a 2×2 contingency table.
p_j = larger of the two marginal proportions in the other variable.
Wherever $p_i = p_j$, the maximal ϕ equals 1.0. To apply this to Table 13.10, C and E,

$$\begin{aligned}\phi_{max} &= \sqrt{\left(\frac{.50}{.50}\right)\left(\frac{.25}{.75}\right)} \\ &= \sqrt{.3333} \\ &= .58\end{aligned}$$

Computations with formula (13.24) are greatly facilitated by use of Table G where values of $\sqrt{p/q}$ and $\sqrt{q/p}$ are given. Formula (13.24) can be broken into the two components $\sqrt{p_j/q_j}$ and $\sqrt{q_i/p_i}$ whose product gives the maximal phi.

Figure 13.4 provides a graphic solution to the same equation for values of p_i from .50 through .98 and p_j throughout the same range. These ranges will take care of all practical situations in which ϕ would ordinarily be computed. It is recommended that the maximal ϕ that suits any given situation be considered when interpreting an obtained ϕ as representing a strength of the *intrinsic* relationship between two variables. The word *intrinsic* is stressed here, because the actual size of ϕ indicates the degree of practical, predictive value of the relationship. Predictive value is actually restricted by inequality of p_i and p_j.

The Coefficient of Contingency.—It has been shown how a ϕ coefficient can be derived from chi square. Phi squared, for a 2×2 table, is equal to chi square divided by N. For this reason ϕ^2 has been called the *mean-square contingency*. By analogy, we might call ϕ the mean contingency, although this name is not used for it. When there are more than two classes in either X or Y, or in both, however, there is another correlation index, called the *coefficient of contingency*, and it is designated by the letter C. The formula for deriving it from chi square is

$$C = \sqrt{\frac{\chi^2}{N + \chi^2}}$$ (Coefficient of contingency computed from chi square) (13.25)

Like ϕ, the coefficient of contingency is restricted in size, but not to the same extent. When the number of categories is large (at least five each way), C approaches the Pearson r in size. If the categorized data

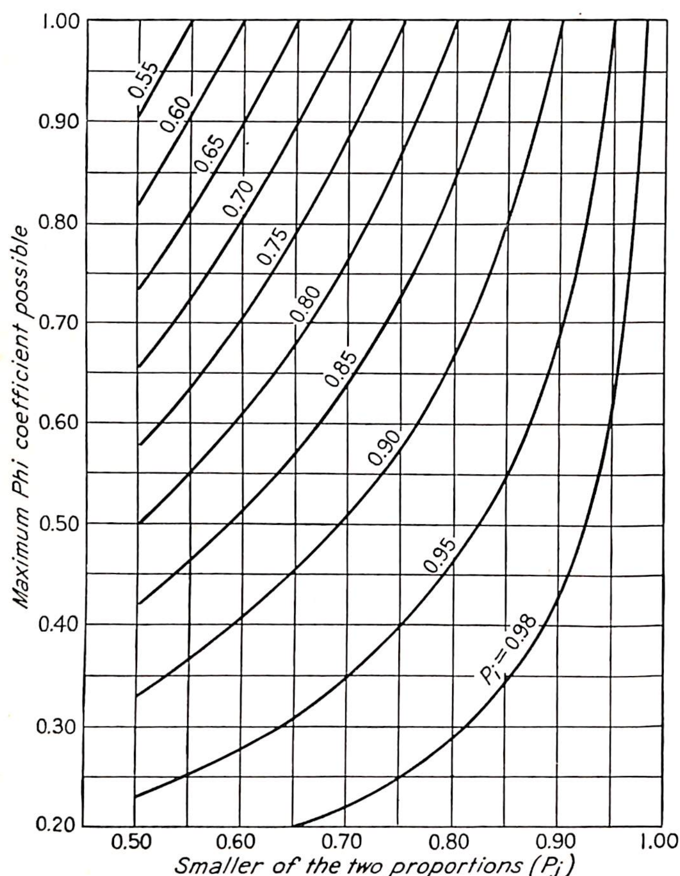


FIG. 13.4.—Maximum phi coefficients for different combinations of proportions of cases in the categories in X and Y having the larger frequencies.

represent continuous, normal distributions, if N is large, and if class intervals are of approximately equal size, the correction procedures applied to the Pearson r, described later in this chapter (Table 13.15), may be applied to the C coefficient. If the data are in genuine categories (point distributions, or nearly so), it is best to interpret C as it is. The maximum C for each given number of categories each way is shown in Table 13.11.

TABLE 13.11.—MAXIMAL VALUES ATTAINABLE FOR A COEFFICIENT OF CONTINGENCY WITH DIFFERENT NUMBERS OF CATEGORIES IN BOTH X AND Y VARIABLES

Number of categories	2	3	4	5	6	7	8	9	10
Maximum C	.707	.816	.866	.894	.913	.926	.935	.943	.949

The standard error of C involves so much computation that it is hardly worth the effort to estimate it. A formula for this is given by Kelley.[1] For testing the hypothesis of zero correlation in a population, the chi square from which C is derived will serve very well.

PARTIAL CORRELATION

The Meaning of Partial Correlation.—A partial correlation between two things is one that nullifies the effects of a third variable (or a number of other variables) upon both the variables being correlated. The correlation between height and weight of boys in a group where age is permitted to vary would be higher than the correlation between height and weight for a group at constant age. The reason is obvious. Because boys are older, they are both heavier and taller. Age is a factor that enhances the strength of correspondence between height and weight. With age held constant, the correlation would still be positive and significant, because at any age taller boys tend to be heavier.

If we wanted to know the correlation between height and weight with the influences of age ruled out, we could, of course, keep samples separated and compute r at each age level. But the partial-correlation technique enables us to accomplish the same result without so fractionating data into homogeneous groups. When only one variable is held constant, we speak of a *first-order partial correlation.* The general formula is

$$r_{12.3} = \frac{r_{12} - r_{13}r_{23}}{\sqrt{(1 - r^2_{13})(1 - r^2_{23})}} \quad \text{(First-order partial coefficient of correlation)} \qquad (13.26)$$

In a group of boys aged twelve to nineteen, the correlation between height and weight (r_{12}) was found to be .78. Between height and age, $r_{13} = .52$. Between weight and age, $r_{23} = .54$. The partial correlation is therefore

$$r_{12.3} = \frac{.78 - (.52)(.54)}{\sqrt{(1 - .52^2)(1 - .54^2)}} = \frac{.4992}{.7189} = .69$$

[1] Kelley, T. L. Statistical method. New York: Macmillan, 1923. P. 269.

With the influences of age upon both height and weight ruled out or nullified, then, the correlation between the two is .69.

As another example with three variables, the correlation between strength and height (r_{41}) in this same group was .58. The correlation between strength and weight (r_{42}) was .72. Although there is a significantly high correlation between strength and height, we wonder whether this is not due to the factor of weight-going-with-height rather than to height itself. So we hold weight constant and ask what the correlation would be then. Will boys of the same weight show any dependence of strength upon height? The correlation is given by

$$r_{41.2} = \frac{.58 - (.72)(.78)}{\sqrt{(1 - .72^2)(1 - .78^2)}}$$
$$= \frac{.0184}{.4343}$$
$$= .042$$

By partialing out weight, it is found that the correlation between height and strength nearly vanishes. We conclude, therefore, that height *as such* has no bearing upon strength, but only by virtue of its association with weight does it show any correlation at all.

Second-order Partials.—When we hold two variables constant at the same time, we call the coefficient a *second-order partial r*. The general formula is

$$r_{12.34} = \frac{r_{12.3} - r_{14.3}r_{24.3}}{\sqrt{(1 - r^2_{14.3})(1 - r^2_{24.3})}} \quad \text{(Second-order partial coefficient of correlation)} \qquad (13.27)$$

In using this formula, the subscripts will have to be modified to suit the choice of variables. Here we are assuming that we want to know the correlation that would occur between X_1 and X_2 with the effects of X_3 and X_4 eliminated from both. It is clear that this formula requires the solution of three partials of the first order previously.

As an example of this partial, we may cite the correlation between strength and age with height and weight held constant. This would mean that if a group of boys having the same height and weight were taken, would older boys be stronger? The raw correlation between age and strength was .29. The second-order partial also turned out to be .29. This means that it seemingly makes no difference whether we allow height and weight to vary or whether we do not; the relation between age and strength is the same within the range examined.

Some Suggestions Concerning Partial Correlation.—Needless to say, unless the assumptions necessary for computing the Pearson r's involved are fulfilled, there is little excuse for using them as the basis for com-

puting partial correlations. There are actually few occasions in psychology and education when a partial r is called for. The partialing out of such things as chronological age is perhaps the most common instance in which it is a useful device. It is not to be recommended as a lazy man's substitute for experimental control and fractionation of data. The newer processes of analysis of variance and tests of significance of statistics from small samples make experimental planning seem more important and the treatment of results more satisfactory without resort to partial correlations. It is inadvisable, in any case, to carry the partial-correlation method much beyond the first-order stage. Beyond this, the structure of the relationships becomes very much involved, and one is bringing more and more raw r's into consideration, each with its own fallibility. The building of an elaborate superstructure of statistics upon foundation stones that are not highly accurate in themselves often leads to questionable results.

Reliability and Significance of an Obtained Partial r.—The standard error of a partial coefficient of correlation is the same as for a Pearson r except that the number of degrees of freedom should be used in the denominator. The general formula is

$$\sigma_{r_{12.34\ldots m}} = \frac{1 - r^2_{12.34\ldots m}}{\sqrt{N - m}} \quad \text{(Standard error of a partial } r) \qquad (13.28)$$

where m is the number of variables involved.

Some Special Problems in Correlation

The Relativity of All Coefficients of Correlation.—It is apparent that the size of the coefficient of correlation depends to some extent upon the method of computing it. What is more important, coefficients computed between the same two variables by the same procedure will vary not only from sample to sample but from population to population. If there are any really absolute correlations in the universe, all variables except the two being held constant, those correlations are probably either zero or 1, or close to either of those values. With contaminating variables left in, the correlations are usually between zero and 1. It is therefore really meaningless to speak of *the* correlation between intelligence and character (if it is assumed even that we know what those variables are and have properly measured them) or even between age and height or any other common variables without at the same time specifying what kind of sample we measured.

A coefficient is always relative to the kind of population sampled and to the manner in which the measurements were made. In reporting coefficients of correlation, any writer should be very careful to state all the pertinent factors that bear upon the size of his obtained correlation coefficients, and

any reader should accept interpretations only when the significant circumstances are kept in mind. A few of the more common sources of variations of size of r will be reviewed briefly in what follows.

The Variability in the Correlated Variables.—The size of r is very much dependent upon the range of ability or, in more general terms, the variability of measured values, in the correlated sample. The greater the variability, the higher will be the correlation, everything else being equal. It should be easier to predict individual differences in scholarship in a class with *IQ*'s ranging from 50 to 150 than in a class where the range is restricted to 90 to 110. If the restriction were to a range of zero (all *IQ*'s being equal) there should be no correlation whatever—the limiting case, in which, of course, no r could be computed at all. Often we know the correlation between some predictive index, such as aptitude-test score and scholarship or some vocational criterion of success as derived from one group of individuals, but we shall be applying the same index to other groups with different ranges of ability, larger or smaller. What will be the effectiveness of predictions in the new groups?

In the selection of personnel by means of tests, as during World War II, research on selective instruments was constantly beset with this very practical problem. New tests were put into use in the selection of personnel, and they correlated substantially with tests that were used in selection. The result was that the men who went into training represented only a higher segment of the population from which selection was to be made by the new tests. The validity of a test could be estimated only for this higher segment of restricted range. And yet, it was the validity in the total population that it was important to know, for it is that validity which indicates the full selective value of the test. The coefficient of validity in the restricted group is almost invariably smaller than what it would be in an unrestricted group.

In a research program such as that on the selection and classification of aviation trainees during World War II, the problem of restriction of range became quite important. Near the end of the war, about 50 per cent of the applicants for aircrew training failed to pass the general qualifying examination, and of these as many as 75 per cent failed to qualify for a particular type of training. Furthermore, it was desired to correlate tests with advanced-training achievement criteria and even combat performance after many more had been eliminated at various stages of training. The proportions of the original applicants who survived to these stages were rather small. Restriction of range was very great.

Karl Pearson, many years ago, provided a solution to this problem that applies under certain conditions. The variables being studied must be

normally distributed in the population and we must know certain parameters or estimates of them in order to solve the problem in any particular situation. We need to know the relation of the dispersions in the restricted and unrestricted populations, either in terms of the variable on which selection occurred or on the basis of some variable correlated with it. We also need to know the correlation in the restricted population between the variable we wish to validate and the criterion of success in training or on the job. There are three formulas of practical use in this problem, each of which recognizes the availability of certain information and the need for validation of a certain kind of variable.

CASE I.—Restriction is produced by selection on the basis of X_1 and there is knowledge of standard deviations in X_1 for both restricted and unrestricted groups. The correlation r_{12} is known in the restricted group. The correlation R_{12} for the unrestricted group is estimated by

$$R_{12} = \frac{r_{12}\left(\frac{\Sigma_1}{\sigma_1}\right)}{\sqrt{1 - r^2_{12} + r^2_{12}\left(\frac{\Sigma^2_1}{\sigma^2_1}\right)}} \quad \text{(Correlation corrected for restriction of range, Case I)} \qquad (13.29)$$

where r_{12} = correlation between X_1 and X_2 in the restricted group.
σ_1 = standard deviation in measurements on X_1 in the restricted group.
Σ_1 = standard deviation in the same variable in the unrestricted group.

In this and in the next two formulas, capital letters stand for values pertaining to the unrestricted population and lower-case letters refer to the restricted population.

The application of this formula is as follows. Suppose that the selection test (X_1) correlated .30 with the training criterion in the group selected on the basis of the test. The standard deviation in the unrestricted group (Σ_1) was 20 and that in the restricted group (σ_1) was 10. The solution is

$$R_{12} = \frac{.30(20/10)}{\sqrt{1 - .09 + (.09)\frac{20^2}{10^2}}}$$

$$= \frac{.60}{\sqrt{1 - .09 + .36}}$$

$$= \frac{60}{\sqrt{1.27}}$$

$$= .53$$

CASE II.—Restriction is produced by selection on the basis of X_1 and there is knowledge of standard deviations for X_2 in both restricted and unrestricted samples and of the correlation r_{12} in the restricted group. The correlation in the unrestricted group is estimated by

$$R_{12} = \sqrt{1 - \frac{\sigma^2_2}{\Sigma^2_2}(1 - r^2_{12})} \quad \text{(Correlation corrected for restriction of range, Case II)} \qquad (13.30)$$

where σ_2 = the standard deviation on X_2 in the restricted group.
Σ_2 = the standard deviation on X_2 in the unrestricted group.

This formula would apply when we correlate two selection tests, when we have selected on the basis of one test (X_1) but know the change of range from knowledge of variances in the other test (X_2). One or both of the "tests" might be a composite score derived from a combination of several tests. An example of this from aviation psychology was the correlation of an experimental test with the pilot stanine (composite aptitude score) when selection had been made on the basis of the stanine and it was more convenient to use the change in dispersion on the test. If we assume the same restricted correlation (r_{12} = .30) as in the previous illustration, also that the restricted and unrestricted standard deviations are 10 and 20, respectively,

$$\begin{aligned} R_{12} &= \sqrt{1 - \frac{10^2}{20^2}(1 - .30^2)} \\ &= \sqrt{1 - \frac{.91}{4}} \\ &= \sqrt{.7725} \\ &= .88 \end{aligned}$$

CASE III.—Restriction is produced by selection on variable X_3, on which variable the restricted and unrestricted standard deviations are known. We wish to estimate the unrestricted correlation R_{12}, when we also know r_{12}, r_{13}, and r_{23}. The formula is

$$R_{12} = \frac{r_{12} + r_{13}r_{23}\left(\frac{\Sigma^2_3}{\sigma^2_3} - 1\right)}{\sqrt{\left[1 + r^2_{13}\left(\frac{\Sigma^2_3}{\sigma^2_3} - 1\right)\right]\left[1 + r^2_{23}\left(\frac{\Sigma^2_3}{\sigma^2_3} - 1\right)\right]}}$$

(Correlation corrected for restriction of range, Case III) (13.31)

where the symbols are defined similarly to those in formulas (13.29) and (13.30). This formula would apply to the correlation of a new, experi-

mental test X_1 with a practical criterion X_2, when selection had been made on the basis of a third variable (pilot stanine, for example) X_3.

The reader may have been somewhat surprised at the rather radical change in correlation that occurred as we corrected for restriction of range in the two hypothetical problems above. To show that these changes are not unreasonable, some data will be cited from the AAF results.[1] An experimental group of more than a thousand pilots had been permitted to enter training without any selection whatever on the basis of either qualifying or classification tests. We know, then, how the pilot stanine and certain classification tests correlated with the graduation-elimination criterion at the end of training. We can also arbitrarily pull out a high segment of the total sample and within that limited sample compute validity coefficients. The results are given in Table 13.12 for

TABLE 13.12.—VALIDITY COEFFICIENTS FOR SELECTIVE TESTS AND A COMPOSITE SCORE FOR THE SELECTION OF PILOT STUDENTS WITH AND WITHOUT RESTRICTION OF RANGE

Variable	Correlation in the total group (N = 1036)	Correlation in the selected highest 13 per cent (N = 136)
Pilot stanine	.64	.18
Mechanical principles	.44	.03
General information	.46	.20
Complex coordination	.40	−.03
Instrument comprehension	.45	.27
Arithmetic reasoning	.27	.18
Finger dexterity	.18	.00

the instance in which a rather high, but not unknown, selection of the top 13 per cent occurred. It can be seen that where there were substantial correlations in the unrestricted sample the correlations in the selected group often shrank close to zero, and in one instance, to a trivial negative r. On the whole, those tests that correlated highest with the stanine lost most in validity correlation because of selection on the basis of the stanine.

Evaluation of the Correction Formulas for Restriction.—It should be repeated that the problem of restriction is important, and that if one

[1] Thorndike, R. L. (Ed.). Research problems and techniques. AAF aviation psychology research program, Report No. 3. Washington, D.C.: Government Printing Office, 1947.

wishes to avoid wrong conclusions, when a substantial amount of selection has been made, one should apply correction procedures. Had we taken the second (restricted) set of coefficients in Table 13.12 seriously, without other knowledge to the contrary, we would probably have concluded that formerly valid tests, and even the stanine, had lost their former validities that were known early in the war when selection was a cause of little restriction.

It should be remembered that the formulas rest on the assumption of normal distributions of the population on the variables used, and the Pearson product-moment r is presupposed. The use of the biserial r or tetrachoric r as an estimate of it raises considerable question when selection is severe. Experience tends to show, however, that when the biserial r is used as the validation coefficient, the formulas tend to underestimate the unrestricted correlation. Better formulas are probably being developed and time will tell whether they will replace those based upon Pearson's solution.[1] The standard errors for these corrected coefficients are unknown, but it is probable that they are much larger than those for Pearson r's of comparable size.

Correlations in Heterogeneous Samples.—Studies of validity of tests and examinations have frequently been faulty from a number of standpoints. The use of school marks as criteria of success in training is in itself a questionable procedure, school marks being derived as they generally are on the basis of measurements of questionable reliability and validity and contaminated with irrelevant factors. This situation alone stacks the cards against high validity coefficients for predictive indices at the start.

There is another factor working against fair tests of validity that we shall face particularly here, a factor also dependent upon the unwarranted faith in school marks as absolute and dependable measures of scholarship. This factor is the indiscriminate pooling of marks from different subjects and from different instructors and treating them as if they were of the same kind of coin. Any cursory inspection of grade distributions in a single institution of learning will show that marks are not by any means of constant value when obtained from different sources. The reader is referred to the situation in Fig. 14.2 where students in an English course making the same score in a common achievement examination were assigned different marks in different sections and by different instructors, probably within the same section. If it is assumed that the comprehensive examination was a valid measure of the students' relative degree of mastery of the objectives of the course, it can be seen how much other

[1] Thorndike, *ibid.*

factors must have entered into the determination of the final mark in the course.

Reference to Fig. 14.2 will show that there is quite a range of scores, from about 85 to 125, within which students were assigned marks all the way from F to B. Only as between marks of F and A is there rather complete lack of overlapping. Striking as this situation is, it is probably rather representative of how much lack of correlation there is between school marks and genuine achievement. Much of this is due to the fluctuation of marking ideas and ideals from instructor to instructor. This variation from set to set of marks when they are collectively correlated with other measures is bound to alter the apparent amount of correlation.

As an example, in six sections of freshmen English, *within* sections the correlation between quiz averages for the semester and a final comprehensive examination ranged from .63 to .92, with an over-all correlation within sections, *when intersection differences had been eliminated*, of .83. Yet when the six sections were combined, *with intersectional differences left in*, the correlation was reduced to .71. It was interesting to find that *between* sections the correlation was −.17, which means that there was a very slight tendency for sections with average lower achievement to be given a higher average quiz mark! This fact accounts for the reduction in correlation from .83 to .71 when sections were combined.[1]

Figure 13.5 pictures the kind of situation just described, in somewhat exaggerated form, in Diagram II. Diagram II is best understood by contrasting it with Diagram I. In the latter we have a homogeneous combination of four subsamples drawn from the same population. The correlation between X and Y within each subsample is indicated by a smaller ellipse. All the ellipses are of about the same shape, indicating about the same degree of correlation of X and Y. The x marks indicate the means of Y and X within each subsample. If we combine the four samples, we obtain a distribution described approximately by the large dotted ellipse. Note that the proportions of the large ellipse are about the same as for each small ellipse, indicating the same level of correlation within the composite distribution as within each subsample. Note, also, that the distribution of the four means forms roughly an ellipse of similar proportions. If the correlation between means of Y and means of X differs from that within subsamples, the correlation of X and Y in the composite sample will differ from that within subsamples.

In Diagram II of Fig. 13.5 we have a very different situation. While

[1] Further discussion of "within" versus "between" correlations when groups are combined will be found in E. F. Lindquist's Statistical analysis in educational research. 219*ff.*

within each subsample the correlation between K and L is the same, the subsamples did not arise from the same population so far as means are concerned. An ellipse drawn to inclose the x's would slant in the direction to assure a negative correlation between means of K and means of L. The effect of this can be seen in the dotted line enclosing all subsamples. Its form suggests approximately zero correlation. Such situations are not uncommon. In general practice, if it is doubtful whether subsamples arose by random sampling from the same population, it would be best to compute correlations within subsamples separately or to apply equivalent procedures which we will not take the space to describe here.[1] The hypothesis of homogeneity of samples can be made by means of t tests or F tests as described in Chs. 9 and 10.

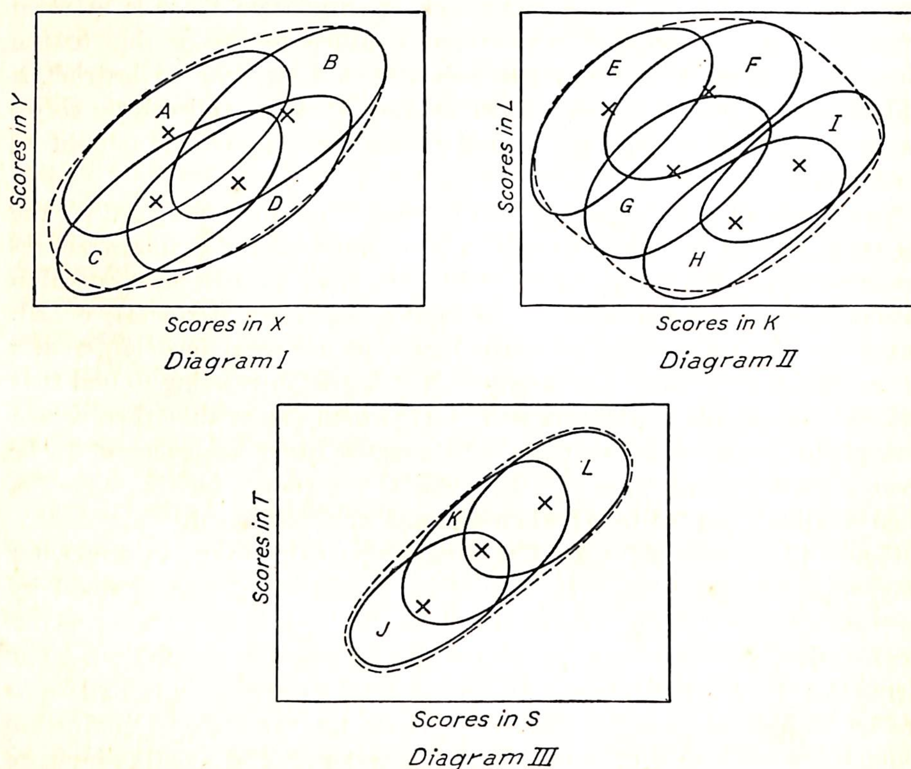


Fig. 13.5.—Illustration of correlations in homogeneous and heterogeneous groups of subsamples.

The Correlation of Averages.—It was stated in an earlier chapter in connection with tests of significance of differences between statistics

[1] See Lindquist, *op. cit.*

(Ch. 9) that the correlation between averages of samples is equal to the correlation between individual pairs of measurements. *This statement assumes random samples from a homogeneous population.* Diagram I in Fig. 13.5 illustrates this kind of situation and shows how an r obtained within one sample can be used as an estimate of a correlation between means. Diagram II shows how a correlation coefficient obtained within a single sample might be very misleading as to the amount of correlation between means. This shows an instance in which the correlation between means is decidedly lower, if not reversed in sign, than that within samples.

The correlation between means could also be higher than that within samples, as Diagram III shows. An example of this would be the correlation between *IQ* and salary. Correlating individuals, we should find some positive correlation, but because of great variations in salary at any single *IQ* value, the correlation might not be very high. If we divided men into sets according to vocation and correlated *average IQ* with *average* salary, the coefficient would probably be very high. This is because people of different *IQ* levels gravitate to certain occupations, and occupations as such have established characteristic salary scales. Other factors that make for individual differences in salary *within* occupations are thus minimized in importance. The sampling is biased the moment we divide groups along occupational lines.

Averaging Coefficients of Correlation.—One solution to the problem of correlations in some heterogeneous samples is to estimate the correlation between X and Y within each subsample and then average the coefficients in order to obtain a single estimate of the population correlation. This would presumably describe the relation between X and Y throughout the composite sample, free from whatever sampling biases there may have been in segregating the subsamples. Before averaging coefficients, however, we must make the assumption that the several r's did arise by random sampling from the same population—same with respect to the degree of correlation. It should go without saying, also, that we have correlated the same variables in all samples. The test of homogeneity of the r's themselves would be based upon their standard errors.

There are several procedures sometimes used in averaging r's. Coefficients of correlation are not values on a scale of equal metric units; they are index numbers. Differences between large r's are actually much greater than those between small r's. If the few sample r's to be averaged, however, are of about the same value and if they are not too large, a simple arithmetic mean will suffice. If the r's differ considerably in size and if they are too large (above .80) some writers urge the procedure using Fisher's $\mathbf{z}$ coefficients. This is illustrated in Table 13.13. It con-

TABLE 13.13.—DEMONSTRATION OF AVERAGING COEFFICIENTS OF CORRELATION WHEN r'S DIFFER IN RANGE AND IN SIZE

Sample *A*		Sample *B*		Sample *C*		Sample *D*	
Mean of r	z method	Mean of r	z method	Mean of r	z method	Mean of r	z method
.45	.48	.75	.97	.35	.37	.65	.78
.50	.55	.80	1.10	.55	.62	.85	1.26
.42	.45	.72	.91	.68	.83	.98	2.30
.38	.40	.68	.83	.50	.55	.80	1.10
.55	.62	.85	1.26	.58	.66	.88	1.38
Σ 2.30	2.50	3.80	4.75	2.66	3.03	4.16	6.82
M_z	.50		1.014		.606		1.364
M_r .46	.46	.76	.77	.532	.543	.832	.877

sists of transforming each r into a corresponding **z** (Table H may be used for this purpose), finding the arithmetic mean of the **z**'s, and finally, converting the mean **z** back to the corresponding r.

The results of Table 13.13 show differences to be expected in the use of an arithmetic mean of r's and of corresponding **z**'s. Samples *A* and *B* have the same range of r's, those in *B* being merely .30 greater than those in *A*. In sample *A*, agreement is perfect in the results from the two methods. In sample *B*, the mean r by the **z** method is .01 higher (.77 as compared with .76). In samples *C* and *D* there is much more spread in the r's averaged. For the r's of moderate size, sample *C*, the **z** method gives a result only .01 greater than the simple mean of r's. In the high coefficients, however, the difference is about .05. There is serious question whether r's differing as much as these would satisfy the belief that they came from the same population by random sampling, and hence would probably not be candidates for averaging. When a few r's do satisfy this belief, the chances are that any discrepancy between a simple mean of r's and an average obtained by the **z** method would be so small as compared with the standard error of r that we could readily forgo the use of the extra effort of the **z** method. If the r's did come from the same population, a mean of several would be a much more reliable estimate of population correlation. With the requirements satisfied, we could add degrees of freedom from the different subsamples to represent the degrees of freedom of the mean r and interpret its reliability and significance accordingly.

Weighting Coefficients in Averaging.—One more requirement should be mentioned, particularly if the last operation, combining degrees of freedom, is to be carried out. That is to weight the obtained r's in averaging

them. The weight for each sample is its number of degrees of freedom ($N - 2$). In using the **z** method, the weights are applied to the **z**'s. A discussion of weighted averages was given in Ch. 4.

The Correlation of Parts with Wholes.—We frequently want to correlate a part measurement, such as a part of a test battery, or a test item, with the whole of which it is a part. Since the variance of the total is in part made up of the variance of the component, that fact alone introduces some degree of positive correlation. The greater the relative contribution to the total variance by the component, the more important is this "spurious" factor. It is possible in a particular instance that the part is totally *uncorrelated* with the remaining parts and yet will be correlated with the total. If it is negatively correlated with the remaining parts, it will be less negatively correlated with the total.

If each part contributes statistically about the same amount of variance to the total or if the part is one of a great many, so that its proportion of contribution is relatively small, we can compare correlations between parts and total with some confidence that they are compared on a very similar basis. But if these conditions do not obtain, we should do better to correlate each part with a composite of all other parts. When such a composite is unknown or is hard to obtain, we can still estimate the correlation by means of the formula

$$r_{pq} = \frac{r_{tp}\sigma_t - \sigma_p}{\sqrt{\sigma^2_t + \sigma^2_p - 2r_{tp}\sigma_t\sigma_p}} \quad \text{(Correlation of part with a remainder, knowing correlation of part with total)} \quad (13.32)$$

where p = part score.
t = total score.
$q = t - p$, in other words, the total with the part excluded.

In the correlation of test items each with the total score of the test of which they are a part, particularly, it is important to know about how much a part would correlate with the total when there is really no relationship at all. We can estimate this, but only under the condition that each part has the same variance and there is zero intercorrelation among all parts. Under these special conditions the average amount of correlation of a part with the total is given by the equation

$$\bar{r}_{pt} = \frac{1}{\sqrt{n}} \quad \text{(Average correlation of a number of parts, of equal variance and zero intercorrelation, with their total)} \quad (13.33)$$

in which n = the number of parts.

In a test composed of 10 such items, the average r_{pt} would equal .316; with 20 items the corresponding figure would be .224; with 30, .183; and with 40, .158. These values should serve as guides. Any obtained part-

whole coefficients of these sizes even though statistically significant might reflect an actual zero correlation of part with total. Application of the correction formula above would be a check upon these low correlations.

If we should want to know the correlation of a part with a whole of which it is a part and we already know the correlation of the part with the remainder of the whole, the estimate is made by the equation

$$r_{pt} = \frac{\sigma_p + r_{pq}\sigma_q}{\sqrt{\sigma^2_p + \sigma^2_q + 2r_{pq}\sigma_p\sigma_q}} \qquad (13.34)$$

(Correlation of part with whole knowing the correlation between part and the remainder)

in which the symbols have the same meaning as in formula (13.32). The utility of this formula is probably rather limited. It is given primarily to show what happens when two parts that correlate zero are combined. If r_{pq} is .0 in formula (13.34), the numerator reduces to σ_p. The denominator is actually the standard deviation of the composite $(p + q)$. The deduction is that if two parts correlate zero, when combined the correlation of the part with the total will be equal to the ratio of the standard deviation of the part to that of the total.

Index Correlation.—This is usually called *spurious index correlation* for the reason that when indices such as *IQ*, *EQ*, or *AQ* are correlated with each other, r is markedly influenced by the fact that these ratios have in common such factors as chronological age and mental age. *IQ*'s from two different tests are derived from the *MA*'s obtained from the two tests *each divided by the same CA*. If there is a range of *CA* in the group correlated, this fact in itself introduces some positive correlation.

Table 13.14 will show by means of a purely fictitious and overdrawn picture how this phenomenon works. For eight children who differ in

TABLE 13.14.—DEMONSTRATION OF HOW INDEX NUMBERS MAY ACQUIRE A HIGH DEGREE OF CORRELATION BECAUSE OF A COMMON DENOMINATOR: AN EXTREME CASE

Child	Chronological age	Mental age I	Mental age II	IQ I	IQ II
A	5.0	7	8	140	160
B	5.5	8	8	145	145
C	6.0	7	7	117	117
D	6.5	8	7	123	108
E	7.5	8	8	106	106
F	8.0	7	8	88	100
G	8.5	8	7	94	82
H	9.0	7	7	78	78

Correlation between mental ages I and II = .00
Correlation between *IQ*'s I and II = .92

chronological age from five to nine inclusive, mental-age ratings on two different tests are given. These are obviously selected children, since their mental-age values hover at seven and eight in a haphazard manner. Note, however, how the *IQ*'s spread, from 140 through 78. The spread in *IQ*'s is almost entirely due to the spread in chronological ages. Since each child has the *same* chronological age for both *IQ*'s, that same denominator of the ratio of his *MA* to *CA* assures that his *IQ*'s will be about the same. Some *IQ*'s go up together in the two tests for children of low *CA* and others go down together, for children with higher *CA*. The correlation computed between *IQ*'s is .92. The same sort of phenomenon goes on in the actual situation to a lesser extent when there is an appreciable range of chronological age.

In the author's opinion, the term *spurious* is not to be confined to this type of situation in particular; for in a sense, all correlations are spurious to the extent that they are influenced by the conditions under which they were obtained. If one remembers what *IQ*'s are and interprets correlations between them accordingly, no particular falsification of the facts is in question. The important thing is that one should correlate variables in the full knowledge of how the measurements were obtained, if possible, and should report to his readers the facts needed for wise interpretation, whether it be variability of the correlated group or range of *CA*'s involved when *IQ*'s have been correlated.

The real difficulty comes when investigator or reader takes *IQ*'s to be some real, absolute properties of individuals, on the one hand, and when someone not oblivious to the common *CA* factor plays it up as a fatal source of "error," on the other hand. Both should remember the relative nature of all correlation coefficients. The important thing is that the wary investigator should not attribute his results to some supposed real nature of psychological or educational phenomena when some property of statistical treatment is really responsible. Nor will the sophisticated critic fail to grant the utility of certain procedures shown to be fruitful under the circumstances of operation even when some "spurious" element has entered the picture. Errors, too, are relative matters. What is an error from the point of view of one frame of reference may be the truth when the frame of reference is changed.

Correction in r for Errors of Grouping.—If in computing a Pearson r by means of grouping data in class intervals, a small number of classes either way has been used, the estimate of correlation is lowered to some degree. In the limiting case, of two classes each way, the computed r is less than two-thirds of the r had there been no grouping. When the number of intervals is 10 both ways, r is about 3 per cent underestimated.

For any number of classes in X or in Y, we can correct for the error of grouping by dividing r by a constant corresponding to that number of classes.

The correction is necessary because errors of grouping yield overestimates of the standard deviations, as was pointed out in Ch. 5. If Sheppard's correction has been applied to both standard deviations, no further correction is necessary in the coefficient of correlation.

Table 13.15 supplies the list of constants given by Peters and Van Voorhis to be used in making corrections in r.[1] Correction is made for

TABLE 13.15.—CORRECTION FACTORS FOR ERRORS OF GROUPING IN THE COMPUTATION OF PEARSON'S r WHEN DISTRIBUTIONS ARE NORMAL AND MIDPOINTS OF INTERVALS STAND FOR CASES IN THE INTERVALS

Number of intervals......	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Correction factor......	.816	.859	.916	.943	.960	.970	.977	.982	.985	.988	.990	.991	.992	.994
Squared correction factor.........	.667	.737	.839	.891	.923	.941	.955	.964	.970	.976	.980	.983	.985	.987

the number of categories or intervals in Y as well as in X. The correction factors are used in the following manner. Suppose we have an obtained r of .61 in a problem with 8 intervals in X and 9 in Y. The correction factors for these numbers of intervals are .977 and .982, respectively. The correction is made by dividing the obtained r by the product of the two correction factors. In terms of a formula,

$$r_c = \frac{r}{c_x c_y} \quad \text{(Coefficient of correlation corrected for coarse grouping)} \quad (13.35)$$

in which c_x and c_y are the correction factors for variables X and Y, respectively, based upon the number of class intervals in each. Applied to the correlation of .61 with 8 and 9 categories in X and Y,

$$r_c = \frac{.61}{(.977)(.982)}$$
$$= .626 \text{ (or .63)}$$

When there are the same number of intervals in both X and Y, the cor-

[1] Peters and Van Voorhis, *op. cit.*, p. 398.

rection factor is the same for both and the factor squared would be called for in the denominator of formula (13.35). The factors squared are given for this purpose in Table 13.15.

When the number of intervals in either X or Y is less than 10 it is good practice to apply this correction procedure; certainly when the number of intervals is 8 or below. There is most to be gained in accuracy of estimate of r when the obtained r is large; little to be gained if r is small, particularly if the sample is small. When the corrective change is small compared with the size of the standard error of r, there is little use in making the correction. It should be remembered that the correction factors given in Table 13.15 are designed especially for the situation in which the midpoint of an interval is the index number for cases in that interval, the intervals are equal in size, and the distributions are normal. For other, less common situations, see the reference below.[1]

Correction of ϕ for Coarse Grouping.—Since the phi coefficient is a product-moment estimate of correlation, the question arises as to whether it is ever subject to this kind of correction. This question should arise only when one or both variables are actually continuously measurable and we want a more realistic estimate of correlation that describes the relationship that exists when the variable is used in graded form. As to number of "intervals" we have two each way when ϕ is computed. The index number for each interval is not the midpoint, however, but is the mean of the cases in the interval. If we can assume actual normal distribution for the variable so correlated, the correction factor is .798. The square of this, which we would use if both variables are normally distributed, is .637. The use of the correction factor .798 would estimate a point-biserial r from ϕ. The use of the squared factor .637 would estimate a Pearson r from ϕ.

An important reservation, however, should be added. Remembering the severe restriction in the size of ϕ, it is probably correct to say that unless $p = p'$, or $p = q'$, in which case only can a maximal ϕ equal 1.0, a corrected phi will not be equivalent to a Pearson r. It is well to remember, too, that some obtained ϕ's are greater than .798 or .637, which would mean corrected coefficients greater than 1.0. In such events, it is probable that the assumptions of continuous, normal distributions are unjustified; the distributions probably are bimodal, if not genuine point distributions. On the whole, the application of correction of ϕ for coarse grouping is so limited and questionable that it is best to think twice before applying it.

[1] Peters and Van Voorhis, *op. cit.*, p. 398.

Exercises

1. Compute by the rank-difference method the correlation between the first 20 scores in any two variables in Data 8*A*. Find the standard error of rho. Interpret your results.

2. Compute for Data 15*A* a correlation ratio for the prediction of Y from X. Find the standard error of eta and the standard error of the estimate. Apply the chi-square test of linearity. Interpret your results.

3. Find from the literature three applications of the correlation ratio. State how the author used eta, and give his reasons, if stated. What subsidiary tests (of linearity, etc.) were made? Make your judgment as to the effectiveness of the uses of eta in the cases cited.

4. If you have mastered the analysis-of-variance procedures as described in Ch. 10, make the application as suggested in this chapter to Data 15*A*, following your solution of the correlation ratio.

5. In the data in Table 13.7, combine the distributions receiving marks of A, B, or C, into a single composite; also, in another composite, combine those receiving marks of D and F. Compute for these data a biserial r between scores and marks. Find the standard error of r_b. Interpret your results.

6. Compute a tetrachoric coefficient of correlation for Data 14*A*. Determine whether or not the correlation is probably significant. If the Thurstone computing diagrams are available, check your solution by this means.

7. Cite some fourfold tables found in this book to which the tetrachoric correlation method should be applied, and cite some others to which it should not be applied.

8. Reduce to a fourfold table preparatory to computing a tetrachoric r the scatter diagram given in Data 15*A*. Do the same for Data 11*A* and Data 8*B*.

9. Find in this volume, or any other source, data to which the phi-coefficient method of correlation may properly be applied. Give reasons.

10. Compute a phi coefficient for Data 11*A*, and make the necessary correction to yield an estimate of the Pearson r. If Exercise 8 has been completed, compare with the r_t found there.

11. Find in the literature examples of coefficients of correlation that might be regarded as spurious from some points of view. How did the author interpret them? How would you interpret them?

12. Apply the correction-for-grouping process to some product-moment coefficient you have obtained or to one you find uncorrected in the literature.

13. The Pearson r for the data in Table 13.7 is .74. Correct it for errors of grouping. How does the change in the corrected r compare with σ_r? How do the uncorrected and corrected Pearson r's compare with the tetrachoric r given for the same data?

14. Determine the following partial r's for Data 16*A* : $r_{34.2}$, $r_{41.2}$, $r_{21.5}$, $r_{51.2}$. Interpret your results. Which of these coefficients have little practical meaning?

15. Determine the following partial r's for Data 16*B*: $r_{31.2}$, $r_{51.4}$, $r_{21.3}$, $r_{45.2}$, $r_{31.24}$. Interpret your results. Suggest other partial r's that might be of importance to know about, and tell why.

CHAPTER 14

PREDICTION OF ATTRIBUTES

One of the most important fruits of scientific investigation and one of the most exacting tests of any hypothesis is the ability to make predictions. So important is this topic that it deserves to have considerable space devoted to it. Particularly is this true for the reason that statistical reasoning is basic to all predictions. Statistical ideas not only guide us in framing statements of a predictive nature but also enable us to say something definite concerning how trustworthy our predictions are—about how much error one should expect in the phenomenon predicted. The practical significance of this cannot be questioned. The significance even for the scientific investigator is too often unrecognized or forgotten.

One can find amateur prognosticators for almost any kind of event on every hand. Little note is made of the success or failure of their predictions. A few successes are sufficient basis for vindication of the "prophet," and many failures are quickly forgiven and forgotten. The old adage "Where ignorance is bliss 'tis folly to be wise" must have been invented to fit this particular situation. On the other hand, the psychologist or educator who falls short of perfect predictions is often immediately condemned and his further predictions thought to be discredited. The average uninformed person is somehow partial to vague and "magical" means of prediction, and he can readily overlook their shortcomings, whereas he will not tolerate the statistically hedged prediction that also yields to him a more exact knowledge of its limitations. If he could only realize how poor the predictions of the amateur prophet actually are, he would perhaps have a more ready respect for the scientific prediction of events in human affairs. It is the purpose of this chapter, and the next two, to illustrate the kinds of predictions the statistically oriented investigator makes and how he not only does not blind his eyes to his failures but brings them clearly into the light.

General Types of Prediction.—Although in this volume we have generally emphasized measurement, we have had to recognize from time to time that complete measurements cannot be made and that data are sometimes obtained as merely classified in categories. The latter type of data we recognize as *enumeration data* rather than as measurements. It

is a matter of assigning attributes to cases rather than quantitative evaluations on a linear scale, for example, identifying individuals as to sex, race, political party, or criminality. Although such data are not allocated to linear-scale positions, we can still make predictions from them and of them from other information. We thus have four cases of predicting:

1. Attributes from other attributes—as when we predict incidence of criminality from sex, race, or religious creed.

2. Attributes from quantitative measurements—as when we predict criminality from scores on tests of ability or of behavior traits.

3. Measurements from attributes—as when we predict probable test scores from sex, socioeconomic status, or martial status.

4. Measurements from other measurements—as when we predict achievement in school from *IQ*-test scores.

General Ways of Evaluating Accuracy of Prediction.—Predictions are obviously sound if they prove to be correct. The degree of correctness is indicated by *how often* or *how nearly* we hit the mark. In the case of predicting attributes, our success can be numerically indicated in terms of the percentages of "hits." But a more accepted way among statisticians is to ask how much better our predictions are than if we had not used the information we have—in other words, if we had not tried to predict one thing from the knowledge of another but merely from a knowledge of the predicted population itself. A more crude way of saying it would be to ask how much better our predictions are than guesswork. But this does not mean *pure* guesswork, as we shall see later.

In predicting measurements, whether from attributes or from other measurements, we ask a similar question. But whereas in predicting attributes for cases, we work in terms of the *number* of hits or misses, in predicting measurements, we work in terms of *how far* on the average we have missed the mark. We compare this average deviation between fact and prediction with the average of the errors we should make without using the knowledge we did as a basis of prediction.

Let us see in a preliminary way what this means. We can predict that a student's mark in a course will be somewhere in the range from A to F inclusive, and most probably it will be a mark of C, which more students earn than any other mark. This prediction is made without knowledge of the student's scholastic-aptitude score, and its margin of error is measurable in terms of the standard deviation of the distribution of marks of all students. If we used knowledge of the students provided by aptitude-test scores, we should predict some to earn marks higher than *C* and some lower than C. The average of our deviations between prediction and fact will now be smaller than the standard deviation of the distribution of all

marks. The difference between these averages of deviations tells us how much the knowledge of aptitude scores has improved our predictions.

Predicting Attributes from Other Attributes

Predictions Can Be Made in Both Directions.—As our first example of prediction of attributes from other attributes, let us consider the data in Table 14.1. Here we have the numbers of persons in a "depressed"

Table 14.1.—Distribution of Responses to the Question, "Would You Rate Yourself as an Impulsive Individual?" as Given by Two Extreme Groups of Students

Group	Response			
	Yes	?	No	Total
Depressed	72	45	133	250
Not depressed	106	35	109	250
Both	178	80	242	500

group who responded by saying "Yes," "?" and "No" to the question, "Would you rate yourself as an impulsive individual?" and also the numbers of a group described as "not depressed." The individuals in these two categories are the highest and lowest quarters of a sample of 1,000 students who were ranked in terms of a provisional scoring on a personality inventory. Table 14.1 provides us with two prediction problems. We can attempt to predict the verbal response to the question, knowing whether the person is in the depressed or not-depressed group; or we can attempt to predict the group to which a person belongs, knowing what response he has made. Let us take the prediction of verbal response first.

The Principle of Maximum Likelihood.—Considering first the depressed group by itself, we find that the largest number of them respond with "No." Taking each member of the depressed group as he came along, we should predict for him the response "No." If all 250 came up for inspection, we should be correct 133 times out of 250, or 53.2 per cent of the time. For other samples from the same depressed population, we should expect a similar ratio of correct predictions. This illustration sets the pattern for all predictions of attributes from attributes. The prediction always observes the *mode* or most frequent attribute in the segment of the population chosen at the moment. For the not-depressed group, the mode is also at the response "No"; hence that is our prediction also

for them, and our percentage of accuracy is 43.6 per cent, not so high as before but higher than if we had predicted either "Yes" or "?" for this group. Such predictions follow the *principle of maximum likelihood* or *maximum probability*. Either a depressed or a not-depressed person in this population is more likely to respond "No" than anything else; so that is our prediction.

The Forecasting Efficiency in Predicting Attributes.—How good are these predictions? Since we have predicted the same response for both depressed and not depressed individuals, we suspect that knowing to which group the person belongs helps us little if any to predict his response. A comparison of the percentages of correct predictions, however, tells us that we can be more sure of our prediction of "No" if the person is depressed than if he is not. But no matter from what group the person comes, our prediction is the same; so it is as if we could make no use of the knowledge of his group affiliation for this purpose.

Let us compare the number of successes of prediction made with and without knowledge of group affiliation. Taking both groups combined, we should predict for each person at random the response "No," and we should be correct 242 times in 500, or 48.4 per cent. In the two groups predicted separately, we found successes of 133 and 109, which combined give us 242 correct hits, or 48.4 per cent. We have thus gained no more accuracy in predicting responses from a knowledge of group affiliation than we could attain without this knowledge. The *forecasting efficiency* in predicting response from knowledge of group is therefore just zero. The work of calculating forecasting efficiency may be seen more clearly if summarized as in Table 14.2.

TABLE 14.2.—PREDICTIONS OF RESPONSE FROM KNOWLEDGE OF THE GROUP MEMBERSHIP

Group membership	Predicted response	Number correct	Per cent correct
Depressed	No	133	53.2
Not depressed	No	109	43.6
Total	...	242	48.4
Correct without knowledge		242	48.4
Excess with knowledge		0	0.0

The second prediction problem here is to reverse matters and predict group membership from knowledge of the response. All persons responding "Yes" we should predict to be members of the not-depressed group,

since 106 actually are, as compared with 72 who are not. Again the *modal* attribute is our prediction. For those responding "?" the prediction is membership in the depressed group, and so also for those responding "No." The percentages of correct predictions are given in Table 14.3 for each response and for all combined. Altogether, there are

TABLE 14.3.—PREDICTIONS OF GROUP MEMBERSHIP FROM KNOWLEDGE OF VERBAL RESPONSE TO THE QUESTION

Response	Predicted group	Number correct	Per cent correct
Yes	Not depressed	106	59.6
?	Depressed	45	56.3
No	Depressed	133	55.0
Total	....	284	56.8
Correct without knowledge		250	50.0
Excess with knowledge		34	13.6

284 correct predictions, or 56.8 per cent. Without knowledge of which response each person made to the question, but with knowledge that half the total population are depressed and half are not, our expected number of chance successes is 250. Our predictions *with* knowledge of responses yielded an excess of 34 or a *forecasting efficiency* of 13.6 per cent. We can say that our predictions with knowledge of response to the question is 13.6 per cent better than those made without this knowledge would be.

Prediction Not Equally Good in the Two Directions.—It is now well apparent that we can predict successfully group membership from knowledge of responses in this problem, whereas we cannot predict response from knowledge of group membership. It is not always true, as it is here, that successful prediction is possible in one direction and *entirely* impossible in the other, but it is a quite common finding that prediction is better in one direction than in the other when two variables are concerned. It will often clarify thinking about predictive problems to keep this fact in mind. It is sometimes assumed by the uninformed that if A can be predicted from B, B can, in turn, be predicted from A. Such an assumption is likely to lead the unwary investigator into logical and practical difficulties when it is seriously wanting in applicability. This is a more serious matter in dealing with attributes than in dealing with measurements, for in the latter case the predictability of one measured trait A from a measured trait B is usually not very divergent from the predictability of B from A.

The Sampling Procedure in Prediction of Attributes.—The evaluations of predictions already given are meaningful and useful. There is still the problem of how significant the decisions based upon the sample may be for the population. This calls for application of sampling statistics. For this purpose we can adapt the use of chi-square, ϕ, and t tests, all of which have been previously described. Their application here contains some new features that need to be explained.

The Cell Square Contingency Method.—We can compute a chi square for the entire contingency table involved in the prediction problem, and that would be meaningful as an over-all index of significance of predictive value somewhere among the categories. As we saw in the previous examination of predictions, however, some predictions are apparently better than others within the same table. By breaking chi square down into components, or rather, by examining the contributions to chi square from the different categories, we obtain a more analytical picture of each one's significant contribution to prediction. Table 14.4 shows the customary steps in the solution of chi square. The last segment of the

TABLE 14.4.—DEMONSTRATION OF THE CELL SQUARE CONTINGENCY METHOD OF TESTING CONTRIBUTIONS TO PREDICTION

Group	Expected frequencies f_e				Discrepancies $f_o - f_e$			
	Yes	?	No	Total	Yes	?	No	Total
Depressed.........	89	40	121	250	−17	+5	+12	0
Not depressed.....	89	40	121	250	+17	−5	−12	0
Both...........	178	80	242	500	0	0	0	0

Group	Squared discrepancies $(f_o - f_e)^2$			Cell square contingencies $\frac{(f_o - f_e)^2}{f_e}$			
	Yes	?	No	Yes	?	No	Total
Depressed...............	289	25	144	3.247	0.625	1.190	5.062
Not depressed...........	289	25	144	3.247	0.625	1.190	5.062
Both..................				6.494	1.250	2.380	10.124
	$C = .14$		t	2.55	1.12	1.51	χ^2

table, in which are given the cell square contingencies, is particularly to be noted.

The chi square for the entire table is equal to 10.12, which, with 2 degrees of freedom, is significant just beyond the 1 per cent level. We next examine each column of the table, for the sum of the cell square contingencies for that column (the column square contingency) indicates the degree of significance to be attached to the category it represents. For the response "Yes," the sum is 6.49. This may be regarded as a chi square for a two-cell table and tests the hypothesis that the depressed and the not-depressed groups should have responded "Yes" in equal frequencies to the question. With one degree of freedom, the departure from the hypothesis is significant almost at the 1 per cent level of confidence. The square root of chi square with one degree of freedom is equal to t, hence t for this response is 2.55. For the other responses, "?" and "No," the t values are 1.12 and 1.54, both insignificant. Thus, we have a decision as to the sampling stability of the gains in accuracy of prediction as given in percentage terms in Table 14.3. Those percentages are 59.6, 56.3, and 55.0 for the three responses, respectively. Only the first seems significant.

As for the prediction of response from knowledge of group membership, the answer lies in the sums of the rows of cell square contingencies in Table 14.4. These sums are the same: 5.06. With 2 degrees of freedom, they fail to be significant at the 5 per cent level. This outcome agrees with the decision based upon Table 12.2, where it was found that there were no excess correct predictions attributable to knowledge of group membership, depressed versus not-depressed. More accurately interpreted, the row sums indicate that the distribution of responses of 250 depressed individuals does not differ significantly from that of the 500 depressed and not-depressed combined. The same may be said for the not-depressed group. When both are considered together, however, their mutual departure from a common, hypothetical distribution (that of the 500 combined) is sufficient to yield a chi square of 10.12, which *is* significant. The corresponding coefficient of contingency (C) equals .14, which is another index of over-all predictive value. Because the chi square from which C was derived is significant at the 1 per cent level, so is C significantly different from zero correlation.

Response Significance as Indicated by Phi.—Another approach, which applies in the special situation in which one of two categories is to be predicted from knowledge of another variable in more than two categories, uses the phi coefficient. Here we are interested only in the prediction of depressed versus not-depressed group membership from knowledge of response to a question. This approach is virtually an example of

the more general application of ϕ to the validation of responses to test items (see Ch. 17). When there are only two alternative responses to an item, the predictive value of the one response equals that of the other response, for lack of more than one degree of freedom. A ϕ coefficient would be quite suitable to indicate the correlation of each response to the item with a two-category criterion. When there are more than two responses, as in the present illustration, we can validate each response separately, although it is, to be sure, just one item, because there is more than one degree of freedom. The validity of any one response, or its correlation with the criterion, does not automatically determine the validities of the others, though, of course, it will have some bearing upon that validity.

The procedure is demonstrated in Table 14.5. There we have three different 2 × 2 contingency tables, one for determining the ϕ for each response. When validating one response we group the others into one

TABLE 14.5.—TESTING THE BASIS OF PREDICTION PROVIDED BY EACH CATEGORY SEPARATELY BY MEANS OF CHI SQUARE AND PHI

Group	Response			Response			Response		
	Yes	? + No	Total	?	Yes + No	Total	No	Yes + ?	Total
Depressed........	72	178	250	45	205	250	133	117	250
Not depressed....	106	144	250	35	215	250	109	141	250
Both..........	178	322	500	80	420	500	242	258	500
	$\chi^2 = 10.08,\ \phi = .142$			$\chi^2 = 1.49,\ \phi = .055$			$\chi^2 = 4.61,\ \phi = .095$		

category. The two categories when validating response "Yes" are responses "Yes" and "Not yes," and so on. The ϕ's for the three responses are .142, .055, and .095, respectively. This is another basis of comparing the effectiveness of the three responses as discriminating between depressed and not-depressed groups. We cannot be very sure that the differences in size of ϕ's are significant, since we do not have standard errors of the ϕ's. We can test the hypothesis of zero correlation, however, by means of the chi squares, which are 10.08, 1.49, and 4.61, respectively. These are to be interpreted as very significant, insignificant, and significant, for responses "Yes," "?" and "No," respectively. These chi squares come in the same rank order as the column square contingencies (see Table 14.4) but they are somewhat larger than the latter.

The differences are to be attributed to a difference in operations. The sum of the three chi squares (10.08 + 1.49 + 4.61) obviously exceeds the sum of the three column square contingencies because each column is included more than once in the three 2 × 2 tables. There is a difference in meaning, also. In computing the phi coefficients, we have asked, "What is the predictive value of a selected response versus all other responses?" If we predict one group membership in this problem from the responses "Yes," we automatically predict the other group membership for all other responses. We find that it paid to group responses "?" and "No" together but it definitely was not so profitable to group any other pairs of responses. The function of the "?" response was much the same as that of the "No" response. This could have been seen in the original table (Table 14.1), in which the directions of differences in frequencies were apparent. It was also apparent in that the same prediction was made from the two responses. The tests of sampling significance bear out those observations. We would obtain as much predictive value by treating responses "?" and "No" as if they were identical as we would by giving them individual weighting, as shown by the fact that when we combine them the chi square (10.08) is about the same as for the entire contingency table (10.12) when the two responses are kept separate. This is also shown by the fact of insignificant ϕ for the fourfold table featuring the "?" response in Table 14.5.

Predicting Attributes from Measurements

We sometimes wish to decide on the basis of known measurements whether an individual should be expected to be in one category, *e.g.*, to have a certain attribute, or whether he should be expected to be in another. Sometimes it is a matter of making placements in different categories in order that the individual may expect a better consequent adjustment or greater satisfaction. Such is the case when we attempt to predict success or failure for persons for whom we know certain test scores. This problem was solved in principle by Guttmann.[1] Here the author will attempt to provide some workable procedures whereby such predictions can be made and their relative accuracy determined.

Critical Points Dividing Distributions.—In Fig. 14.1, we have two populations, differing in mean, standard deviation, and in N. We wish to find a score on the scale of measurement that will give us the maximum accuracy of prediction, so that we may say of an individual whose score is higher than that point that he is probably a member of the upper

[1] The prediction of personal adjustment. New York: Social Science Research Council, 1941. Pp. 271*ff*.

group and of an individual whose score is lower than that point that he is probably in the lower group and in so predicting, make the minimum number of mistakes. Let us call that critical point E.

According to Guttmann's solution, point E comes on the scale where the two distributions have equal ordinates—in other words, where the two curves intersect (see Fig. 14.1). At this point, persons with scores of this value are equally likely to be members of either group. Above this point, at any score there is greater likelihood that the person belongs in the upper group than that he belongs in the lower group. Below this point, at any score, there is a greater likelihood that the person belongs in the lower group. The terms *upper* and *lower* here apply only to relative

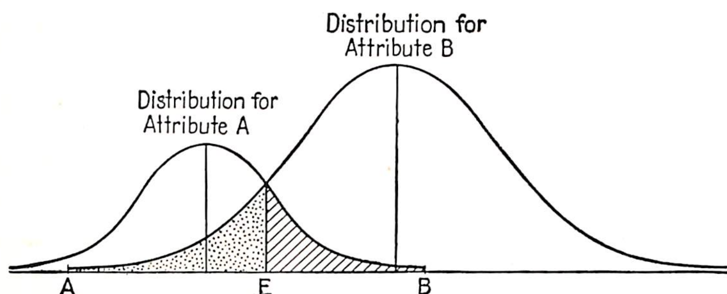


FIG. 14.1.—Distribution of two hypothetical groups possessing two distinguished attributes, A and B, when measured on the same scale of some other variable. The aim is to predict for each person his attribute from knowledge of his score. For those with scores above point E we predict attribute B as being more likely; for those below E, we predict attribute A.

position on the measuring scale. The two distributions are divided according to two qualities or attributes, and it is possession of those attributes that we are trying to predict. As we proceed along above point E, the probability that we are correct in our prediction increases, since the ratio of individuals having attribute B to the number having attribute A keeps increasing. At point B, which is the upper limit of the range of the A group, and above B, we should have absolute certainty of prediction so far as these particular populations are concerned. Likewise, below point A, where the upper distribution ends, we should be absolutely certain that no case possesses attribute B. But if the two populations are taken as wholes, the shaded portions stand for the proportions of individuals incorrectly predicted. The cross-hatched section (of distribution A) represents the A's wrongly predicted to be B's, and the stippled section (of distribution B) represents the B's wrongly predicted to be A's. All the B's above point E are correctly predicted. It is on the basis of these numbers of correctly and incorrectly predicted

cases that we can judge the forecasting efficiency, as we shall see later. First, let us see how point E can be determined.

Locating a Critical Point for an Artificial Dichotomy.—The principle upon which the point of division is made on the continuous variable is a variation of the principle of maximum likelihood. For scores above the critical value, the probability of a case being in the upper category is greater than .5. For scores below the critical value, the probability of a case being in the upper category is less than .5.

The location of the critical division point depends to some extent upon whether the dichotomy is a genuine one or whether it is an artificial one based upon continuous measurements. There are several methods that can be used to solve the problem. Some apply to either kind of dichotomy, some to one or the other but not to both. We will begin with methods that apply to the artificial dichotomy.

As illustrative material, let us use the data in Table 14.6. A large group of students were given the same comprehensive final examination in freshman English. Each instructor was at liberty to use the scores in this examination along with other measurements as he saw fit in deriving a final mark in the course for his students. Taking all marks collectively, for all students receiving a mark of F, a frequency distribution of their examination scores was set up. The same was done for students receiving marks of D, C, B, and A. These are the five distributions listed in Table 14.6 and shown graphically in Fig. 14.2. The amount of overlapping in ability as represented by examination scores among these five groups is noteworthy, but it probably represents a not unusual situation where marks are determined in the customary manner. However that may be, let us say that students receiving F's are, in the judgment of the teachers, failing students, and those receiving D's are D students, etc. These five categories represent five attributes as judged by these instructors. Let us take as our problem the task of predicting what attribute will be assigned to students making certain scores in the examination.

Graphic Methods of Locating the Critical Point.—When the overlapping distributions are plotted as in Fig. 14.2, if they are fairly regular in contour, one can immediately locate the points at which the two distributions intersect. Distributions for attributes F and D intersect just below a score of 60; more exactly, by inspection, at 57 or 58. In this approach, it would be well to locate the point between two whole numbers, because scores are obtained in whole numbers. In this case, we should predict an F for students making a score of 57 or lower, and a mark of D for those making a score of 58 or above (at least up to the critical point between D and C). Between D and C, the critical point, by inspection,

TABLE 14.6.—DISTRIBUTIONS OF SCORES IN A GENERAL ENGLISH EXAMINATION MADE BY STUDENTS RECEIVING VARIOUS MARKS IN THE COURSE

Scores	*A*	*B*	*C*	*D*	*F*
180–189	1				
170–179	1	1			
160–169	5	7	1		
150–159	7	13	3		
140–149	2	26	10	1	
130–139	2	34	24	5	1
120–129	0	40	39	7	0
110–119	1	21	81	13	3
100–109		19	89	28	4
90– 99		4	81	29	9
80– 89		1	42	46	8
70– 79			16	29	11
60– 69			5	20	9
50– 59				6	11
40– 49				1	5
30– 39					3
20– 29					0
10– 19					0
0– 9					1
Sums.......	19	166	391	185	65

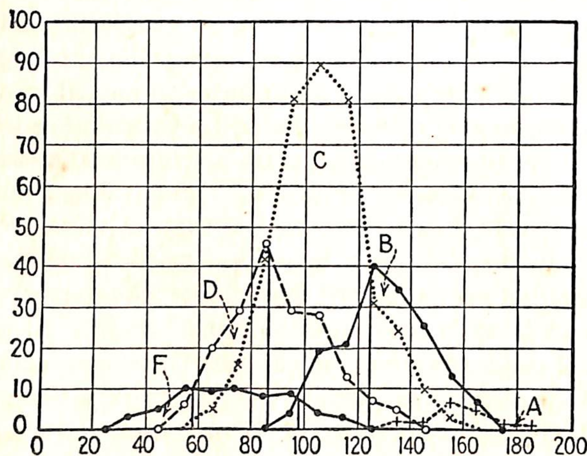


FIG. 14.2.—Distributions for students receiving marks of A to F in freshman English, of scores received in a common final examination.

seems to be at about 87, probably on the lower side. Thus, for scores 58 through 86, we should predict a mark of D. The next critical point seems to come between 124 and 125. The prediction of a C arises for scores 87 through 124. The critical point between B and A is almost impossible to determine but seems to lie in the region of 170 to 175. The small number of A's makes any solution of this kind uncertain.

Should overlapping distributions be irregular in contour, particularly in the neighborhood of the intersection point, if the data are not too limited, and if the smoothing required is rather obvious, it would be well to resort to smoothing before the point of intersection is sought (see Ch. 3 for a description of smoothing procedures).

This graphic method of determining a critical dividing score point may do for rough estimates when samples are large and contours of distribution curves are regular. A better graphic procedure will be described next. It is not only rather useful in practical situations but demonstrates a more general conception of the prediction problem.

Preparatory to the application of this method, the frequency distributions of Table 14.6 were combined in various ways as shown in Table 14.7. In this method we are interested in finding out from the data the probability that an individual who earned a score of a certain size will be in the upper of two groups. In column (1) we have the total composite distribution. In column (2) we have the distribution of only those who received a mark of A. The probability of a student in any class interval on the examination receiving a mark of A is indicated by the proportion of all those in that interval who actually did receive a mark of A. This is an empirical probability, derived from the sample data. We use it as an estimate of the population probability. Not until we go down the column of frequencies in column (2) to the interval 160–169 do we find frequencies of a size that would give us much confidence in the accuracy of the proportion derived from them. In that interval, 5 out of 13 received an A, or a proportion of .38. In the interval 150–159, 7 out of 23, or 30 per cent, received an A. The other columns of the table represent other division points as to upper and lower marking categories. In columns (6) and (7) we are interested in the proportions in the class intervals receiving a mark of C or above.

Figure 14.3 shows graphically the relation between these proportions and the various score levels. The midpoint of each interval is used to represent that interval. This figure demonstrates that the increase in probability of being in an upper of two categories on another variable (marks) is of an S-shaped form with different degrees of skewness. The skewness is related to the over-all proportion in the upper category and

TABLE 14.7.—FREQUENCY DISTRIBUTIONS OF ENGLISH EXAMINATION SCORES FOR STUDENTS RECEIVING MARKS ABOVE CERTAIN DIVISION POINTS; ALSO PROPORTIONS IN EACH UPPER CATEGORY AT DIFFERENT SCORE LEVELS

Scores	(1) f_t	(2) f_a	(3) p_a	(4) f_{ab}	(5) p_{ab}	(6) f_{abc}	(7) p_{abc}	(8) f_{abcd}	(9) p_{abcd}
180–189	1	1	(1.00)	1	(1.00)	1	(.100)	1	(1.00)
170–179	2	1	(.50)	2	(1.00)	2	(.100)	2	(1.00)
160–169	13	5	.38	12	.92	13	1.00	13	1.00
150–159	23	7	.30	20	.87	23	1.00	23	1.00
140–149	39	2	.05	28	.72	38	.97	39	1.00
130–139	66	2	.03	36	.545	60	.91	65	.985
120–129	86	0	.00	40	.465	79	.92	86	1.00
110–119	119	1	.01	22	.185	103	.87	116	.975
100–109	140	0	.00	19	.14	108	.77	136	.97
90– 99	123	0	.00	4	.03	85	.69	114	.93
80– 89	97			1	.01	43	.44	89	.92
70– 79	56			0	.00	16	.29	45	.80
60– 69	34			0	.00	5	.15	25	.735
50– 59	17					0	.00	6	.35
40– 49	6					0	.00	1	(.17)
30– 39	3							0	.00
20– 29	0							0	.00
10– 19	0								
0– 9	1								

f_t = frequency in distribution of all students combined.
f_a = frequency in distribution of students receiving a mark of A.
p_a = proportion of students in each score interval who received a mark of A. Proportions in parentheses are very uncertain owing to the extremely small samples from which they are computed.
f_{ab} = frequency in distribution of students receiving marks of A and B.

to the skewness of the total distribution. With large numbers in the upper category the skewness tends to be positive and with small numbers the skewness tends to be negative. The points are sufficiently in line that one can draw continuous curves through them by inspection (which has been done in Fig. 14.3), except at the tails of some of them where data are incomplete.

While we are interested primarily in the score level at which the probability of an individual's being in the upper category is exactly .5, it is important to note that these functions tell us much more than that. They tell the probability at each score level of an individual's being in

the upper category. We can say that for a score of 120 there is apparently no chance of a student's receiving an A, there are about 31 chances in 100 of his receiving a B or above (with no chance of an A, this amounts to the odds for receiving a B), and there are about 89 chances in 100 of his receiving a C or better. There is possibly one chance in 100 of his failing the course. A student with a score of 70, however, has apparently no chance of receiving an A, or B, about 22 chances in 100 of receiving a C or better, about 77 chances in 100 of receiving a D or better, and conversely, 23 chances in 100 of failing.

To determine the scores corresponding to proportions of .5, by this graphic solution the division points appear to be: between A and B, a

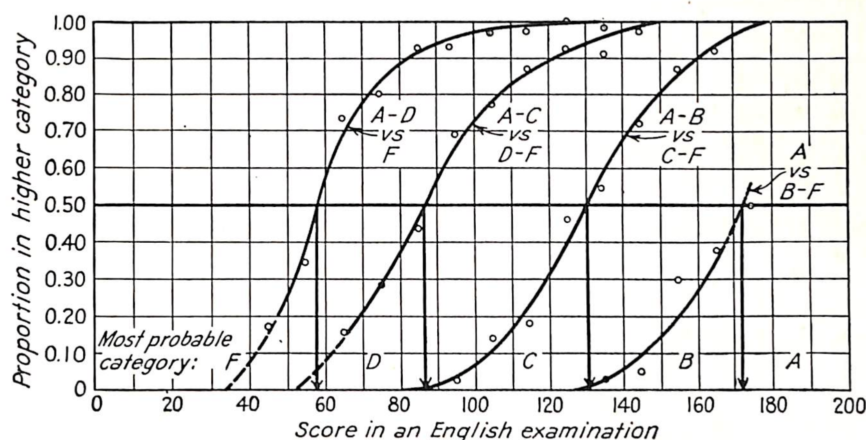


FIG. 14.3.—Proportion of the students who are in each higher letter-grade category at each score level in a common freshman-English examination.

score point between 171 and 172; between B and C, a score point between 130 and 131; between C and D, a score point between 86 and 87; and between D and F, a score point between 57 and 58. The last two coincide with those read from Fig. 14.2. The first is more accurately determined, though still rather uncertain. The estimate of a division of 130.5 between marks B and C differs considerably from the 124.5 that was read from Fig. 14.2. These comparisons alone tell us nothing about the accuracy of either method, except that they agree very closely (within 1 unit) on two and roughly on a third, with intolerable disagreement on the fourth.

Before leaving the two graphic methods, it should be pointed out that a very important difference exists between them. In the first of the two, only two adjacent distributions are considered in determining the critical score that is to separate them. In the second, we consider all cases within the one letter-category distribution and all others above as being in the

upper group and we consider all cases within the neighboring letter-category distribution and all others below as being in the lower group. This kind of problem comes up only when there are several division points to be established; more often there are only two. In the latter instance, all of the distribution in X is involved, just as it is in the second graphic method and as it is in the computational method to follow. Not only does the second graphic method provide more stable values to work with because of larger subsamples but it also follows better statistical principles as expressed in the development of the computational method.

A Computation of the Critical Score.—It has been demonstrated recently that for this type of problem—predicting membership in one of two artificial dichotomies—a formula may be used to estimate the critical score.[1] We must assume for this purpose that both the distributions (in X and in Y) are actually continuous and normal. The formula is

$$X_c = M_x + \left(\frac{zy}{pq}\right)\left(\frac{\sigma^2_x}{M_p - M_q}\right) \quad \text{(A critical division point for maximal accuracy of separation into two categories in a correlated variable)} \quad (14.1)$$

where M_x = mean of the entire distribution, for those in the two categories combined.

p = proportion of the total population in the category having the higher mean score on X.

$q = 1 - p$.

y = ordinate in the unit normal distribution at the point of division of the area under the normal curve with p proportion above it.

z = standard measure of the point at which the division just referred to occurs.

This normal distribution stands for the dichotomized variable in the same manner as it does in connection with the computation of a biserial r. In fact, there is a close relationship between formula (14.1) and the formula for computing a biserial r (formula 13.7). There is an alternative formula for estimating the critical score:

$$X_c = M_x + \left(\frac{zy}{p}\right)\left(\frac{\sigma^2_x}{M_p - M_x}\right) \quad \text{(Alternative estimation of a critical division point)} \quad (14.2)$$

The latter version of the formula is applied to the computation of X_c in the English-examination problem, with the work shown in Table 14.8.

[1] This method was developed by the author and W. B. Michael and its derivation is described elsewhere: Guilford, J. P., and Michael, W. B. The prediction of categories from measurements. Beverly Hills, Calif.: Sheridan Supply Co., 1949.

The four division points by calculation are 167.82, 130.2, 86.5, and 53.1. The second and third are within one unit of those found by the second graphic method. These findings, though very limited, suggest that the second graphic method may be superior to the first and that neither is very satisfactory unless there are a sufficient number of points on both sides of the .5 level to establish the proper location of the curve in the region of that important level. The labor involved in computation of X_c by formula is probably no greater than that for the graphic methods and leaves nothing to guesswork. The graphic method does have one advantage, that it does not require any assumption about the distributions on the two variables.

Accuracy of Predicting Artificial Categories.—The evaluations of predictions of categories when they are made from measurements can be made in a manner similar to those previously described. Our interest may be in the numbers and percentages of correct predictions (or in the numbers and kinds of errors) and in the gain in accuracy of prediction from the new knowledge possessed.

As an illustration, let us take the example of the English-examination data as related to course marks. To note the accuracy of prediction in two categories only, we may use the division between the B students and above and the C students or below. The indications are that the best separation on the score scale should

TABLE 14.8.—WORKTABLE FOR THE COMPUTATION OF THE DIVISION POINTS ON THE SCORE SCALE OF THE ENGLISH EXAMINATION AS COMPUTED BY FORMULA

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)
Upper group	N	M_p	p	z	y	zy	$\frac{zy}{p}$	$M_p - M_x$	(V) $\frac{\sigma^2_x}{(M_p - M_x)}$	$\frac{Vzy}{p}$	X_c
A	19	154.50	.0230	+1.9954	.0545	+.10875	+4.72827	48.95	13.1673	+62.27	167.82
AB	185	132.07	.2240	+0.7588	.2922	+.22703	+1.01353	26.52	24.3039	+24.63	130.2
ABC	576	114.38	.6973	−0.5167	.3491	−.18038	−0.25868	8.83	72.9943	−18.88	86.5
$ABCD$	761	108.32	.9213	−1.4139	.1468	−.20756	−0.22529	2.77	232.6860	−52.42	53.1

$M_x = 105.55 \quad \sigma^2_x = 644.54$

be between a score of 130 and one of 131. It is not possible to make an exact separation of the cases given in grouped form in Table 14.6, since the dividing score point comes within an interval. For the sake of applying the test of goodness of prediction, however, let us assume that the 66 students are evenly distributed over the range 130–139, and that one-tenth of them would have a score of 130. This means about 7 students, 4 of whom are in the A–B mark group and 3 of whom are in the C–D–F group. With these arbitrary, but minor, adjustments, we can arrange the entire sample of 826 students in a 2 × 2 distribution as in Table 14.9.

TABLE 14.9.—SUMMARY OF CORRECT AND INCORRECT PREDICTIONS OF LETTER MARKS A AND B VERSUS C, D, AND F, IN FRESHMEN ENGLISH FROM AN EXAMINATION SCORE

Examination Score

Marks	Above 130	130 or below	All scores
A, B.	95	42	137
C, D, F.	90	599	689
All marks.	185	641	826

Score group	Prediction	Number correct	Per cent correct	Per cent in total group
Above 130.	A or B	95	51.4	16.6
130 or below.	C, D, or F	599	93.4	83.4
Total.	. . . .	694	84.0	

There are several ways of interpreting this table. We can note that there were 132 errors of prediction. If we are interested in predicting marks from scores, with the division point adopted we would wrongly elect 90 to receive marks of A or B and we would wrongly designate 42 to receive marks of C, D, or F. In predicting the 185 who according to high scores should receive A or B we would be correct in 51.4 per cent of the cases. This does not seem very high accuracy, unless we compare it with the proportion of those with A and B marks in the entire group, which is 137/826, or about 16.6 per cent. In predicting the 641 to receive C or below, the accuracy of 93.4 seems very high until we realize that about 84 per cent of the entire sample received similar marks. In comparing the percentages of correct predictions with the percentages of corresponding types of cases in the entire sample, we are going in the direction

of the chi-square test, in which divergency of distribution in the row or columns from the distribution in the marginal frequencies is the indication of departure from a random situation. A more interpretable index of the *degree* of divergence is the phi coefficient. In this problem, chi square is 208.11, which is far above required significance levels. From this we find ϕ to be .50, which indicates the amount of correlation between marks and examination scores when both are dichotomized and used in that manner for prediction purposes.

We could test the accuracy of prediction in similar ways for each of the other division points. The fourfold tables of frequencies would tell their own stories and ϕ would summarize the agreement between prediction and fact. The ϕ might vary somewhat from one division to another. In a multiple-category problem like this one, some might prefer to consider all five mark categories together and note, for each division point, how many errors in predicting marks are one-place errors, how many are two-place errors, and so on. A two-place error, for example, would be predicting a B when a D was obtained. A 5×5 contingency table might be set up with the four critical scores as the division points between categories in variable X. In so far as the widths of categories on the score scale differ, a contingency coefficient, C, would be the summarizing index of correlation to use.

The kind of study of errors of prediction will depend upon what information the investigator hopes to gain from the results. Whenever a procedure depending upon the counting of cases is used, it should be emphasized that rather large samples are needed for dependable comparisons.

Locating a Critical Point in Predicting a Genuine Dichotomy.—When the dichotomy is genuine, the graphic methods that were previously described apply. The division is at the point of equal likelihood, and the graphic methods satisfy that principle for the sample. Assuming that the sample is representative of the population, approximately the same division point should be effective in making predictions in the population.

An example of data that may be treated as a genuine dichotomy is given in Table 14.10.[1] The two categories are "alcoholics" and "non-alcoholics" defined in the clinical sense. The alcoholics were recognized by responsible agencies as problem drinkers. It can be argued that there is a continuum of degrees of tendency toward alcoholism, but clinically and administratively there is a rather definite categorization which divides

[1] These data were adapted from a doctoral dissertation by M. P. Manson, A psychoneurotic differentiation between alcoholics and non-alcoholics. *Quart. J. Stud. Alcohol.*, 1948, **9**, 175–206.

TABLE 14.10.—DISTRIBUTION OF ALCOHOLICS AND NONALCOHOLICS FOR SCORES ON AN ADJUSTMENT INVENTORY

Scores in the inventory	(1) Frequency distributions: Non-alcoholics	(2) Alcoholics	(3) Both	(4) Proportion alcoholic	(5) Percentage distributions: Non-alcoholics	(6) Alcoholics	(7) Both	(8) Proportion alcoholic
66–71	0	1	1	(1.00)	0	0.5	0.5	(1.00)
60–65	0	6	6	(1.00)	0	3.0	3.0	(1.00)
54–59	1	13	14	.93	0.7	6.4	7.1	.90
48–53	1	13	14	.93	0.7	6.4	7.1	.90
42–47	3	17	20	.85	2.2	8.4	10.6	.79
36–41	3	33	36	.92	2.2	16.3	18.5	.88
30–35	2	32	34	.94	1.4	15.8	17.2	.92
24–29	9	32	41	.78	6.6	15.8	22.4	.705
18–23	16	23	39	.59	11.7	11.4	23.1	.49
12–17	36	24	60	.40	26.3	11.9	38.2	.31
6–11	43	7	50	.14	31.4	3.5	34.9	.10
0– 5	23	1	24	.04	16.8	0.5	17.3	.03
N	137	202	339	.596	100.0	99.9	199.9	
M	14.11	32.83	25.27		14.08	32.80	23.44	
σ	10.41	13.93	15.61				15.45	

the two. When in doubt about continuity it is best to treat a dichotomy as being real.

Inspection of the distributions in the table shows that the possibilities for prediction are quite promising. The first graphic method, based upon overlapping of the two frequency-distribution curves, with or without smoothing, gives a division point between scores 18 and 19. For any score of 19 and above we would expect to find more than half of the individuals in this sample alcoholic and for a score of 18 and below less than half alcoholic. The second graphic method gives the same result as the first.

Before accepting this solution as the one we want, however, it is necessary to consider a new aspect to the prediction problem when we are dealing with qualitative categories. Second thought about the alcoholism data will suggest the idea that the distributions as given represent the general population of men very poorly. In the general population, the proportion of alcoholics is extremely small; certainly not 60 per cent, as the data in question show. The data were obviously not selected on the

basis of stratification. In fact, for the purpose of the investigation, contrasting groups of about equal size were desired. Suppose that we had alcoholics represented in line with their proportion in the general population. When we came to apply the first graphic method, with relatively much smaller frequencies in that group, the intersection of the curve with that for the nonalcoholic group would have been at a much higher score, if indeed it intersected at all. By the second graphic method, the proportions of alcoholics might have been less than .5 at all score levels. No solution by the principle of equal likelihood would then have been possible. Another type of solution is therefore called for; one less dependent upon

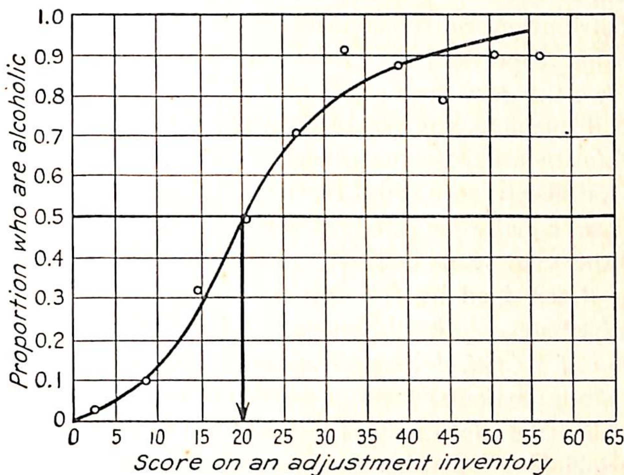


FIG. 14.4.—Proportion of alcoholics at each score level on an adjustment inventory. The problem is to find that score point above which more than half have the property alcoholic.

the proportions of the two kinds of individuals in the general population, if the principle of equal likelihood is to be applied.

Assuming that we have qualitative categories, and that we are attempting to predict one quality or another, it would seem logical to treat the two as being of equal importance. In the data of Table 14.10 we may regard the mean of 14.11 as being characteristic of nonalcoholics as a species, also the form of distribution they gave. This is true if there was no biasing of sampling *within* this group as such. Likewise, we may regard the distribution of scores for alcoholics as characteristic of *their* population. This suggests a solution which would allow the two species equal representation. To achieve equal representation we may convert the obtained frequencies into percentage frequencies. These appear in columns (5) and (6) of Table 14.10. Beside them, in column (7), are

given the sums of percentage frequencies in the different class intervals, and in column (8) are given the proportion of alcoholics at each score level. The graphic solution based upon these is shown in Fig. 14.4, which yields a critical division point between scores 20 and 21. Following this approach we may say that with scores 21 and above the odds are greater than .5 that the individuals have the property of alcoholism and with scores of 20 and below the odds are less than .5 for this property. We will consider later how many and what kind of errors this division point would entail.

When the two category groups are equated for size, as in the method just described, a much simpler solution is possible in certain situations. If the two distributions on the continuous variable are both symmetrical and of the same dispersion, the critical point will be at the unweighted mean of the two category means (M_p and M_q). This would be true, also, if with equal dispersions any positive skewness in the one distribution is compensated for by a like degree of negative skewness in the other. If all one wants is a division score and if these conditions are satisfied, the mean of the two means equally weighted will serve. For the data on alcoholism, the mean of the two means is 23.44. This is somewhat higher than the critical point determined by the graphic method, because the two distributions differ markedly in dispersion and in skewness.

Computation of a Critical Value Dividing Genuine Dichotomies.—Without assuming any particular form of distribution for the continuous variable except that it be continuous, a critical value that will approximately satisfy the principle of equal likelihood may be estimated by the formula[1]

$$X_c = M_x + \left(\frac{.5 - p}{pq}\right)\left(\frac{\sigma^2_x}{M_p - M_q}\right) \quad \text{(Critical value on } X \text{ dividing cases into most probable categories)} \qquad (14.3)$$

where M_x = mean of all X values.

p = proportion of the cases in the category having the higher mean of X values.

$q = 1 - p$.

M_p = mean of X values for category higher on X.

M_q = mean of X values for category lower on X.

σ^2_x = variance in the total distribution on X.

Let us apply this formula to the prediction of sex membership of high-school students from knowledge of hand-grip scores. For a sample of 171 boys and 246 girls, the two means (M_p and M_q) were 37.35 and 20.68, respectively. The mean of all cases combined was 27.51. The variance

[1] From Guilford and Michael, *op. cit.*

of the combined group was 115.38. The proportions (p and q) were .410 and .590. Applying formula (14.3),

$$
\begin{aligned}
X_c &= 27.51 + \left[\frac{.5 - .41}{(.41)(.59)}\right]\left[\frac{115.38}{37.35 - 20.68}\right] \\
&= 27.51 + \left(\frac{.09}{.2419}\right)\left(\frac{115.38}{16.67}\right) \\
&= 27.51 + (.37205)(6.9214) \\
&= 27.41 + 2.58 \\
&= 30.09
\end{aligned}
$$

This result tells us that students earning a score of 31 or above are more likely to be boys than girls; those with scores of 30 or below are more likely to be girls.

An alternative formula requires less information. It reads

$$X_c = M_x + \left(\frac{.5 - p}{p}\right)\left(\frac{\sigma^2{}_x}{M_p - M_x}\right) \qquad \text{(Alternate to 14.3)} \quad (14.4)$$

where the symbols are as defined previously. While this formula is more convenient in computing, formula (14.3) is somewhat more meaningful. It will pay to examine (14.3) to see what may be expected as p varies and as $M_p - M_q$ varies.

First, note that the critical score is the mean of all the X values plus an increment. This increment is positive and X_c will be above the general mean when p is less than .5. It will be negative and X_c will be below the general mean when p is greater than .5. The division of cases in making predictions is in the same direction as that in the population. When $p = .5$, the increment becomes zero and the critical value equals M_x. This fact is true regardless of the amount of correlation existing between X and the categories. When p deviates very far from .5, the ratio becomes quite large and likewise the increment. The critical value may even go outside the distribution, which would mean that we would predict all cases to be within the category having the greater frequency. If 90 per cent of a population, let us say, are in the upper category, X_c might go very low on the scale. If we predicted all, or nearly all, of the cases to be in the upper category, we would, of course, make a very small number of errors.

It is of interest to consider the relation of the increment to the amount of correlation between X and Y. The type of correlation appropriate here is the point biserial. The point-biserial r is proportional to $M_p - M_q$ and inversely proportional to σ_x. This being true, it appears that the

increment is inversely proportional to the amount of correlation. The higher the correlation, the nearer X_c is to the general mean, M_x. When the correlation is perfect, predictions should ordinarily be perfect. For predictions to be perfect, the position of X_c should be such that the proportion expected in the upper category coincides with p, the obtained proportion. As the correlation approaches zero, the critical value departs more and more from M_x and assures the prediction of more and more cases in the more populous category. As r_{pbi} becomes zero, if p does not equal .5, the increment becomes very large and most predictions fall in the more populous group, if not all. Thus, the prediction is determined relatively more by knowledge of X when the correlation is large and by the knowledge of which category is more populous relatively more when the correlation is small, as we should expect.

When Population Proportions Differ from Sample Proportions.—Formulas (14.3) and (14.4) presuppose that the sample proportion is a good estimate of the population proportion. Application of the principle of equal likelihood depends upon this. In the case of the prediction of alcoholism from inventory scores, however, we know the population proportion of alcoholics is very far from the .596 that prevailed in the sample. In the general nonhospitalized population, the proportion might be less than 1 per cent. In a prison population or a hospital population, it would undoubtedly be greater than 1 per cent. In a psychopathic ward it would probably be even greater. How, then, should we apply the formulas? Will we want to observe the principle of equal likelihood under all situations? We saw some doubt cast on its application earlier. Let us apply formula (14.4) to the data on alcoholism, assuming different population proportions for alcoholic addiction; proportions of .333 (one-third), .2, .1, and .01, as well as the .596 of the Manson study and the special case of $p = .5$. We do not have data derived from such populations, but if we assume that the means and standard deviations already found for the two categories of persons hold for the general situation, we can estimate M_x and σ^2_x for populations made up of the specified proportions. The data are given in Table 14.11.

For the obtained proportion of .596 for alcoholics, the X_c which would give the maximal number of correct classifications is 20.08. For an assumed proportion of .50, X_c is 23.47, which is equal to M_x when the two classes are equal in size. This differs from the value estimated by the graphic method in Fig. 14.4, which was approximately 20:3. The two may be expected to coincide, as was suggested previously, when the two distributions have equal dispersions and skewness. They do not satisfy this condition here. If alcoholics made up a third of the popu-

TABLE 14.11.—ESTIMATION OF CRITICAL DIVISION SCORES FOR PREDICTING ALCOHOLISM AS POPULATION PROPORTIONS OF ALCOHOLICS ARE ALLOWED TO VARY

p	M_x	σ^2_x	$M_p - M_x$	$\frac{\sigma^2_x}{M_p - M_x}$ (V)	$\frac{.5 - p}{p}$ (W)	$V \times W$	X_c
.596	25.28	243.77	7.55	32.287	− 0.161	− 5.20	20.08
.500	23.47	237.78	9.36	25.511	.000	0.00	23.47
.333	20.35	214.78	12.48	17.210	+ 0.500	+ 8.60	28.95
.200	17.85	181.56	14.98	12.120	+ 1.500	+ 18.18	36.03
.100	15.98	148.47	16.85	8.811	+ 4.000	+ 35.25	51.23
.010	14.30	112.69	18.53	6.082	+49.000	+297.99	312.29

lation in which predictions are made, the X_c should be at 28.95. If they made up only 1 per cent of the population, it would take a critical score of 312 to find the two kinds of individuals equally represented. This is, of course, well outside the practical range of scores.

It is true that as the proportion of nonalcoholics increases, for the same critical score, 23, for example, the greater the numbers and percentages of mistakes (of the kind diagnosing nonalcoholics as alcoholics) that would be made. To reduce the number of mistakes one would move X_c upward, as the results in Table 14.11 demonstrate. For practical use of the predictive instrument, however, one would have to desert the principle of equal likelihood. Decisions then should be made taking into consideration the relative seriousness of the two kinds of errors. The principle of equal likelihood carries the implicit assumption that the two kinds of error are of equal importance.

Effectiveness of Predictions in Genuine Dichotomies.—The goodness of prediction of the type being discussed here can be evaluated in much the same manner as for the prediction of artificial categories. This is true, particularly, when there are stable and meaningful population proportions in the two categories. In view of the several qualifications mentioned above, however, the kind of evaluation will have to be adapted to fit the situation and to give the most meaningful and pertinent conclusion. The point-biserial r is a general index of correlation that applies here. It will not give the kind of answer often desired in this connection. With a given critical value chosen for X, we have a fourfold contingency table, to which other tests, as described before, apply.

Exercises

1. Using Data 14*A*, make predictions in both directions. Determine the percentages of correct predictions with and without knowledge of categories and the percentage of forecasting efficiency. Discuss the results, including the usefulness of the predictions.

2. Using Data 14*B*, make predictions of whether a student will report "Yes," "?" or "No" to the question about talking when he makes similar responses to the question about walking in his sleep, and vice versa. What are the percentages of accuracy in these various predictions and in the over-all set of predictions?

3. Apply the cell square contingency test to Data 14*B*, testing predictions from different sources. Make any combinations of categories that seem necessary. Compute chi square for the entire table. Draw conclusions.

4. Find a critical total score which will subdivide the total group in Table 13.4 into the most probable categories (passing and failing). Use two graphic methods and a solution by formula. Discuss any discrepancies that may occur.

5. Find a critical division point between boys and girls for the data in Fig. 15.1, which will make the best prediction of sex membership from knowledge of weight. Use formulas (14.3) and (14.4). Evaluate the results of prediction in any way that seems most informative to you.

DATA 14*A*.—RELATIONSHIP BETWEEN FAILING IN COLLEGE AND BEING ABOVE OR BELOW THE MEDIAN IN HIGH-SCHOOL GRADUATING CLASS

Status in high-school class	Failing in one or more courses	No failures in first semester	Total
Above the median	37	340	377
Below the median	49	71	120
Total	86	411	497

DATA 14*B*.—RELATIONSHIP BETWEEN WALKING IN ONE'S SLEEP AND TALKING IN ONE'S SLEEP AS REPORTED BY 1,787 STUDENTS*

Talk in your sleep?	Walk in your sleep?			
	Yes	?	No	Total
Yes	88	9	400	497
?	3	14	194	211
No	7	3	1,069	1,079
Total	98	26	1,663	1,787

* Jenness, A. F., and Jorgensen, A. P. Ratings of vividness of imagery in the waking state compared with reports of somnambulism. *Amer. J. Psychol.*, 1941, **54**, 253–259. Reproduced with the permission of the editor of *Amer. J. Psychol.*

CHAPTER 15

PREDICTION OF MEASUREMENTS

Predicting Measurements from Attributes

The Principle of Least Squares.—What would be the most accurate prediction of the weight of a sixteen-year-old youth? By "most accurate" we mean a weight that, if chosen to predict the weight of each sixteen-year-old selected at random from a certain population, would be closer to the facts in the long run than any other estimate would be. To state the matter in another way, we want a predicted weight that would give us the smallest average discrepancy from the actual weights. For every person, we should find the difference between his actual weight and our prediction in order to obtain the single discrepancy.

Statisticians have good reason to deal here in terms of the *squares* of the discrepancies rather than in terms of the discrepancies themselves. They demand a predicted measurement from which the sum of the squared discrepancies is a minimum. The prediction that will satisfy this requirement has been proven to be the mean of the distribution. In choosing the mean as our prediction, we are following the *principle of least squares.* Whereas in predicting attributes we chose the *mode* of a distribution as the indicator that would give us the smallest *percentage* of error of placement of cases, in predicting measurements, we choose the *mean* as the indicator, which gives us the smallest set of squared deviations from the predicted value.

Predictions Apply to Selected Populations.—In answering the question with which we started this discussion, the best prediction of the weight of a sixteen-year-old, any better knowledge being lacking, is the mean weight of the population of which he is a member. If we wanted this to cover *all* sixteen-year-olds, we should see to it that our distribution from which we derive our mean is made up of a large sample in which both sexes, all races, and all socioeconomic and geographic groups are proportionately represented. We might, however, confine the question to sixteen-year-olds from the United States. We might further confine it to high-school youths in one city, or, even further, to one particular high school. Whatever our restriction in population, the predicted weight will apply only (except by chance) to that kind of population. In fact,

strictly speaking, it will apply only to the measured sample. Whenever we extend our predictions to samples beyond our known population, we always do so at the risk of enlarging errors of prediction.

Errors of Prediction Measured by the Standard Deviation.—In a certain high school in a certain American city, a random sample of 51 sixteen-year-olds had weights distributed as shown in Fig. 15.1. For the sake of an illustration, we shall adopt the sixteen-year-olds in this high school as our population. What we say concerning predictions within this group will hold by analogy to larger, more inclusive populations. The mean of the 51 students' weights is 61.9 kg. and the standard deviation is 13.2.

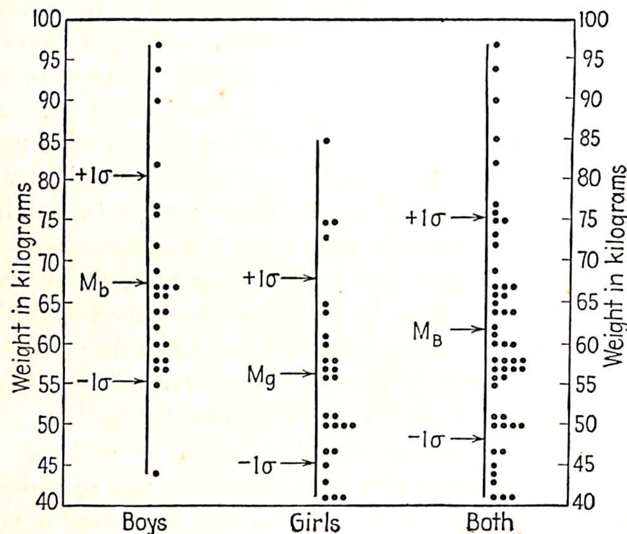


FIG. 15.1.—Distributions of sixteen-year-old high-school boys and girls for weight in kilograms. Each dot represents one individual.

If now the 51 students were listed in alphabetical order and without seeing them we used merely the knowledge of the mean, we should most nearly predict the actual weights if we wrote after each student's name "61.9 kg." The odds are about 2 to 1, as the interpretation of σ goes, that our errors would be no greater than 13.2 kg. either way from the predicted weight. The σ of 13.2 kg. may therefore be taken to measure our margin of error in predicting single cases within the sample, when prediction is based only upon knowledge of the mean.

Any other prediction we might make for all the individuals would yield a larger margin of error, according to the principle of least squares. We should not be very proud of our accuracy of prediction in this instance, and for practical purposes of making decisions for individuals where their

weights are important factors, we should be seriously in error in many cases. But we could do less well in predicting the individuals' weights if we did not even possess the knowledge of their mean. Even if we knew the mean of sixteen-year-olds in general and used that as our predictive value, we should do worse than we did, unless the mean of this small population coincides with that of all sixteen-year-olds. In other words, by knowing one attribute of our population—a group in one American high school—and the mean that goes with that attribute, we reduce the error of prediction to some extent.

Predicting Weight from Knowledge of Sex.—Of the 51 cases in the population of sixteen-year-olds, 24 were boys and 27 were girls. Will it help to predict more accurately if we know each individual's sex? It should, since there is a sex difference in weights. Though many girls are heavier than many boys, the averages are distinctly apart—67.8 for the boys and 56.6 for the girls. Using the attribute of sex to contribute toward the prediction of individual cases and following the principle of least squares, for each boy who came along we should predict his weight to be 67.8 kg., and for each girl, the prediction would be 56.6 kg.

How much will predictions now be improved? The margin of error of predictions for boys is given by the σ of their distribution, which is 12.6 kg., and the margin of error for the girls is given by a σ of 11.3. From this information, we see that both boys' and girls' weights are more accurately predicted than before (when the margin of error was 13.2) and that the girls' predicted weights are more free from error than are the boys'.

As a matter of consistency with previous procedures, let us ask what the percentage of reduction in error of prediction is. For the boys, the change of .6 in the σ is 4.5 per cent, and for the girls, the change in σ is 1.9, or 14.4 per cent.

The Standard Error of Estimate.—There is a way of summarizing the margin of error for all cases combined. This requires the computation of a *standard error of estimate*. It is a kind of summary of all the squared discrepancies of actual measurements from the predicted measurements. In terms of a formula, the standard error of estimate is

$$\sigma_{yx} = \sqrt{\frac{\Sigma(Y - Y')^2}{N}} \quad \text{(Standard error of estimate)} \qquad (15.1)$$

where Y = measured value of a case we are trying to predict.
Y' = predicted value for the case.
N = total number of cases predicted.

The subscript in σ_{yx} tells us that we are predicting variable Y from variable X. In the illustrative problem, Y is the variable of weight, and X is

the variable of sex difference. The sum of the discrepancies squared (see Table 15.1) is 7,288.1; so

$$\sigma^2_{yx} = \frac{7.288.1}{51} = 142.90$$

$$\sigma_{yx} = 11.9$$

The standard error of the estimate, in predicting weight on the basis of knowledge of sex, is 11.9. Using only the knowledge that this is a particular group of sixteen-year-olds with a mean of 61.9, the error of estimate was given by a standard deviation of 13.2. The margin of error using the information supplied by sex difference is 90.2 per cent as large as that without using this information. The reduction in size of error of prediction is 9.8 per cent, which is rather small but represents some gain.

In computing the standard error of estimate in this kind of problem, it is probably more natural to do so by finding the σ's of the two part distributions separately and then combining them. They cannot be combined directly by simple addition or averaging. It is the squared deviations in the two groups that must be combined. The sum of the squared deviations in each distribution can be found by the formula[1]

$$\Sigma x^2_a = N_a \sigma^2_a \quad \text{(Sum of square of discrepancies within one distribution)} \qquad (15.2)$$

where Σx^2_a = sum of the squared discrepancies between prediction and fact (or between measurements and the mean) in distribution A (one of the attribute distributions).

N_a = number of cases in distribution A.

σ_a = standard deviation of distribution A.

When these sums of squared deviations are obtained from all component distributions (distributions A, B, and C, etc.), they may be combined by simple addition to give $\Sigma(Y - Y')^2$. In other words

$$\Sigma(Y - Y')^2 = \Sigma N_k \sigma^2_k \quad \text{(Sum of squares of discrepancies in all distributions)} \qquad (15.3)$$

where N_k = number of cases in any component distribution (distributions A, B, C, etc., in turn).

σ_k = standard deviation of the same distribution.[2]

The work of computing $\Sigma(Y - Y')^2$ for the problem on weights of sixteen-year-olds may be summarized as in Table 15.1. From here the computation of σ_{yx} is exactly the same as previously demonstrated.

[1] *Cf* formula (5.8).

[2] It will be recognized that $\Sigma(Y - Y')^2$ is essentially a sum of squares from which the *within* variance would be computed in analysis of variance (see Ch. 10).

TABLE 15.1.—SUMMARY OF THE COMBINATIONS OF SUMS OF SQUARES FROM DIFFERENT SUBSAMPLES

Distribution	N_k	σ	σ^2	$N_k\sigma^2$
Boys.	24	12.65	160.02	3,840.48
Girls.	27	11.30	127.69	3,447.63
				7,288.11
				$\Sigma(Y - Y')^2$

Other Predictive Indices May Be Introduced.—It should be added that other attributes may be brought into the predictive picture. For instance, if different glandular constitution has a definite bearing on body weight, for example, thyroid functioning, we could subdivide each sex group into two or three categories as to glandular condition. The mean of each new subgroup would then become the prediction for members of that group. The deviations of actual weights from these means would be smaller and the new standard error of estimate would be reduced in size.

If we were successful in singling out all the significant factors correlated with weight and could predict from all of them at the same time, theoretically we could reduce errors of prediction to approximately zero. We can probably never know what all the significant factors are from which weight can be determined, and if we did it might be impossible to assign all the attributes to each individual. We are here speaking of the hypothetical limiting case. Any improvement in predictions approaches that limit. From a practical standpoint, it is always a question of whether the trouble of uncovering and using new descriptive attributes is justified by the gains in predictive accuracy that result.

Estimation of Errors of Prediction in the Population.—The standard error of estimate computed for the weight-prediction problem, strictly speaking, applies to the sample only. It is a biased estimate of the margin of error that would occur in making predictions beyond this particular sample but in the same population. To estimate the standard error of estimate for the population, we need, as usual, to consider degrees of freedom, unless the sample is large. The formula would be the same as (15.1) with the substitution of $N - m$ for N, where m is the number of categories predicted from.

$$_c\sigma_{yx} = \sqrt{\frac{(Y - Y')^2}{N - m}} \quad \text{(Standard error of estimate corrected for bias)} \qquad (15.4)$$

With this formula applied instead of formula (15.1), the corrected stand-

ard error of estimate is 12.2 rather than 11.9. The corrected one is the more realistic one to use in making predictions outside the sample.

Predicting Measurements from Other Measurements

When both known and predicted variables are measured on linear scales and there is some relation between them so that predictions are possible, we have a much more complicated problem. A complete treatment of it involves correlation methods, regression equations, and other procedures.

The Correlation Diagram.—Our illustration of this kind of problem consists of two achievement examinations in a course on educational measurements. In Table 15.2, we have the two distributions grouped in class intervals and the measurements in each class interval broken down to form a distribution of its own in the other test. The class intervals for test X are listed along the top of Table 15.2 and the class intervals for test Y are listed along the left margin.

Prediction of Y from X.—As usual, we have here a double prediction problem; the prediction of a score in Y from a known score in X, and vice versa. Let us consider the prediction of Y from X first. For the individuals in any class interval in test X, the best prediction is the mean of the Y distribution in that column, in other words, the mean of the column (M_c). For each column of Table 15.2, its mean is listed in the

Table 15.2.—Predicting Scores in One Test from Known Scores in Another Test

Test Y	Test X										
	60–64	65–69	70–74	75–79	80–84	85–89	90–94	95–99	f_y	M_{row}	σ_{row}
135–139								1	1	97.0	—*
130–134				1	1	0	1		3	83.7	6.61
125–129				1	0	2	1		4	85.8	5.45
120–124			1	4	4	6	2		17	83.2	5.67
115–119			7	5	7	2	1		22	78.6	5.72
110–114	1	4	2	9	4	2			22	75.9	6.56
105–109	1	1	2	5	1				10	74.0	5.56
100–104	1	3	0	1	1				6	70.3	6.87
95– 99		2							2	67.0	0.00
f_x	3	10	12	26	18	12	5	1	87 = N		
M_c	107.0	105.5	114.9	114.5	116.4	120.3	124.0	137.0			
σ_c	4.08	5.52	4.31	6.83	6.43	4.71	5.10	—*			

* The standard deviation of this array is indeterminate.

next to last row. For the first column, M_c is 107.0. Any person receiving a score from 60 to 64 inclusive in test X will most probably earn a score of

107.0 in test Y. The other means of the columns are similarly interpreted. It will be noticed that there is a general upward trend in the M_c's as we go up the scale in test X, though there are two inversions. In view of the small numbers of cases upon which these means are based, some inversions are not surprising.

The margin of error in predicting Y from X in each column is indicated by the standard deviation of that column. The σ_c's are listed in the last row of Table 15.2. They remain fairly constant, but the range is from 4.08 to 6.83. The significance of the variations in σ_c could be examined by making F tests (see Ch. 9).

The entire picture of predictions and their margins of errors within columns is shown graphically in Fig. 15.2. The circlets show the positions of the column means, and the vertical lines running through them extend from $-1\sigma_c$ to $+1\sigma_c$. In each column, we expect two-thirds of the observed scores to lie within the limits of these lines.

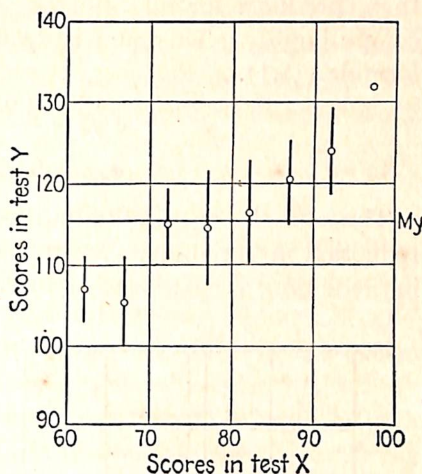


FIG. 15.2.—A chart showing the most probable score in test Y corresponding to each midpoint score in test X, also the range between minus and plus one standard deviation within each column.

Standard Error of Estimate.—In order to obtain a single indicator of the goodness of the prediction of Y scores from X scores, we may compute a standard error of estimate as we did before when predicting measurements from attributes. The work is best organized as in Table 15.3.

TABLE 15.3.—COMPUTATIONS OF THE STANDARD ERROR OF ESTIMATE OF Y SCORES FROM X SCORES

N_c	σ^2_c	$N_c\sigma^2_c$
3	16.67	50.01
10	30.45	304.50
12	18.58	222.96
26	46.63	1212.38
18	41.36	744.48
12	22.22	266.64
5	26.00	130.00
Σ		2930.97 $\Sigma(Y - Y')^2$

For every column, we list first N_c, the number of cases in that column. Second, we list σ^2_c, the squared σ of the distribution in that column. Next we find the product of these two values for that column. The sum of these products for all columns yields $\Sigma(Y - Y')^2$, which we need for computing σ_{yx}. This sum is 2,930.97. From here on the work follows formula (15.1).

$$\sigma^2_{yx} = \frac{2{,}930.97}{87} = 33.6893$$

$$\sigma_{yx} = 5.80$$

The σ of the entire distribution of Y scores is 7.85, so that there is a reduction in variability of 2.05, or 26.1 per cent, a marked improvement in prediction, as such tests go. We may say that the forecasting efficiency for predicting Y scores from X scores as we did is approximately 26 per cent.

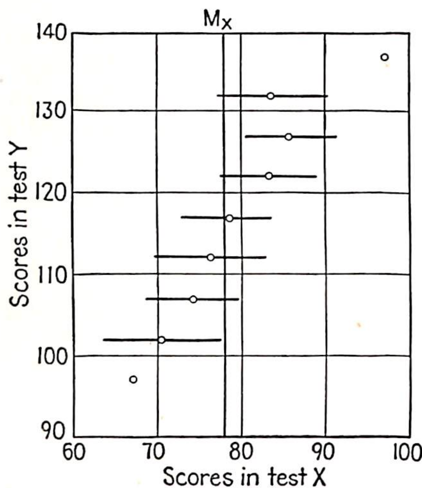


Fig. 15.3.—A chart showing the most probable score in test X for each mid-point score in test Y, also the range between minus and plus one standard deviation within each row.

Predicting X from Y.—The predictions of X from Y are listed in Table 15.2 under M_{row} in the next to the last column. The most probable X score for any interval of Y scores is the mean of the row. The margin of error of the predictions is given in each case by σ_{row}, and these appear in the last column of Table 15.2. To complete the picture of these predictions and their σ's, Fig. 15.3 is presented. The standard error of estimate of the X scores, σ_{xy} (note the order of x and y in the subscript), is equal to 5.93. Since the total σ of the X scores is 7.60, the reduction in error of prediction is 1.67, which is 22.0 per cent. The forecasting efficiency in predicting X from Y is in this problem somewhat lower than the forecasting efficiency (26.1 per cent) in predicting Y from X.[1]

The procedure for predictions by using means of columns and rows is not used very much in practice. It was emphasized here because of the principles it illustrates; principles that underlie the regression methods to

[1] The σ's of the arrays were computed here without applying Sheppard's correction. Had this correction been used, the σ's would have been smaller and consequently σ_{yx} and σ_{xy} would have been smaller.

be described next. The reader will find that the main principle for making predictions of measurements still holds—the principle of least squares. He will also find that the principles for testing accuracy of prediction—the standard error of estimate and the percentage of reduction of errors—also still apply. New ways of estimating them will be shown and their relation to the coefficient of correlation will be explained. In addition, new ways of interpreting the usefulness of predictions will be demonstrated.

Regression Equations

The Meaning of a Regression Equation.—The main use of a regression equation is to predict the most likely measurement in one variable from the known measurement in another. If the correlation between Y and X were perfect (with a coefficient of $+1.00$ or -1.00), we could make predictions of Y from X or of X from Y with maximum accuracy; the errors of prediction would be zero. If the correlation were zero, predictions would be futile. Between these two limits, predictions are possible with varying degrees of accuracy. The higher the correlation the greater the accuracy of prediction and the smaller the errors of prediction.

When we use the means of columns of a scatter diagram as the most probable corresponding Y values, we are actually predicting Y's only for the midpoints of intervals on X, or stated in another way, we are predicting the same Y value for a certain range of values on X. If we have any desire to be more accurate than that, we should like to be able to make predictions for *all* values of X. This the regression line and the regression equation enable us to do.

We found (see Figs. 15.2 and 15.3) that the means of the columns (and of the rows) tended to lie along a straight line, with some minor deviations from strict linearity. We shall now assume that the best predictions of Y from X lie along a line that best fits the means of the columns when those means are weighted according to the number of cases represented in each one. This is known as the *line of best fit*, or the *regression line*. When predicting X from Y, we have another such line for the regression of X on Y. The two regression lines for the achievement-test data will be found pictured in Fig. 15.4. Only when a correlation is perfect will the two lines coincide throughout their lengths. The higher the correlation, plus or minus, the closer together they tend to lie. All such pairs of regression lines intersect at the point representing the means of Y and X; in this case, they cross at $X = 78.15$ and $Y = 115.28$.

The Regression Equations and Regression Coefficients.—From elementary algebra, the student should remember that the equation for a straight line, in general form, is $Y = a + bX$. Such an equation com-

pletely describes a line when a and b are known; they are the *regression coefficients* and must be obtained from the data we have. Leaving out of account for the moment the coefficient a, we should have $Y = bX$, or Y equals b times X. We see from this that b is a ratio, and *it tells us how many units Y is increasing for every increase of one unit in X.* If b were 2, then for every unit of increase in X, Y increases two units. If $b = 0.5$, then for every unit increase in X, Y increases a half unit. The b coefficient gives us the *slope* of the regression line, and it depends upon

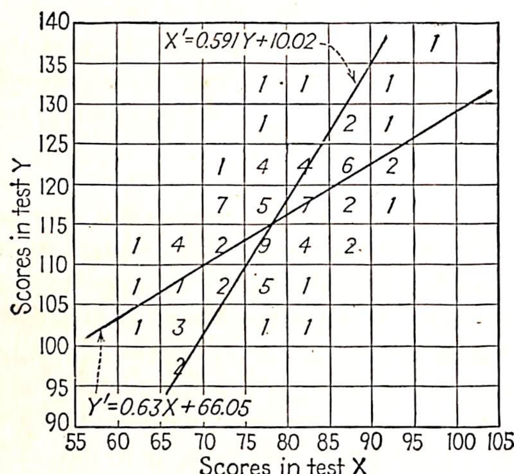


FIG. 15.4.—A scatter diagram for two examinations, with two regression lines represented and their equations.

the coefficient of correlation and the two standard deviations, as in the formula

$$b_{yx} = r_{yx}\left(\frac{\sigma_y}{\sigma_x}\right) \qquad \text{(Coefficient for linear regression of } Y \text{ on } X) \qquad (15.5)$$

where b_{yx}, with the subscripts in that order, implies that we are predicting Y from X, and where this is also true for r_{yx}.[1]

When we want to predict X from Y, we have a different regression equation with a different b, which is given by the formula

$$b_{xy} = r_{xy}\left(\frac{\sigma_x}{\sigma_y}\right) \qquad \text{(Coefficient for linear regression of } X \text{ on } Y) \qquad (15.6)$$

The coefficient of correlation is, of course, numerically the same in both cases, since $r_{yx} = r_{xy}$. But in each case, the b's are different and are equal to r times the ratio of the standard deviation of the *predicted* vari-

[1] For a derivation of formulas for finding regression coefficients, see Appendix A.

able to that of the variable *predicted from*. We frequently speak of the predicted variable as the *dependent* variable and of the one predicted from as the *independent variable*. The reason for this is that in predicting Y from X, we arbitrarily take any value of X that we wish at the moment, whereas the Y we predict from it *is* dependent upon what X we have chosen. Once we have picked out a certain X, Y is immediately fixed by our regression equation.

The regression coefficient a is merely a constant that we must always add in order to assure that the mean of the predictions will equal the mean of the obtained values. As b_{yx} determines the *slope* of the line, a_{yx} determines the general *level* of the line. It is given by the formulas

$$a_{yx} = M_y - (M_x)b_{yx} \qquad (15.7a)$$
$$a_{xy} = M_x - (M_y)b_{xy} \qquad (15.7b)$$

(The a coefficient in a linear regression equation)

where the first one concerns the equation for the regression of Y on X and the second concerns the equation for the regression of X on Y.

The derivation of the entire regression equation is more often accomplished by one composite formula, combining the derivations of a and b into one operation as follows:

$$Y' = r\left(\frac{\sigma_y}{\sigma_x}\right)(X - M_x) + M_y \qquad (15.8a)$$

(Complete statement of linear regression equations)

$$X' = r\left(\frac{\sigma_x}{\sigma_y}\right)(Y - M_y) + M_x \qquad (15.8b)$$

We use Y' and X' here rather than Y and X to show that they are predicted rather than obtained values. Predictions and obtained values rarely coincide unless correlations are nearly perfect.

Applied to the data of Table 15.2, we have

$$\begin{aligned} Y' &= .61\left(\frac{7.85}{7.60}\right)(X - 78.15) + 115.28 \\ &= (.61)(1.03)(X - 78.15) + 115.28 \\ &= .630X - 49.23 + 115.28 \\ &= .630X + 66.05 \\ X' &= .61\left(\frac{7.60}{7.85}\right)(Y - 115.28) + 78.15 \\ &= .591Y + 10.02 \end{aligned}$$

Interpreting these equations, we may say that Y' increases .630 units for every unit increase in X and that X' increases .591 units for every

unit increase in Y. One way of checking the accuracy of the solution of regression equations is to substitute M_x in the first one to see whether Y' is the mean of the Y's and to substitute M_y in the second to see whether we obtain M_x as our prediction of X.

Another check as to the accuracy of computation of the b coefficients is the equation

$$b_{yx}b_{xy} = r^2 \qquad \text{(Relation of regression coefficients to } r^2\text{)} \qquad (15.9)$$

In other words, the product of the two b coefficients is equal to the square of the coefficient of correlation. In this instance

$$(.630)(.591) = .3723 = .61^2$$

The Concept of Regression.—It may help in understanding the regression equations as given in formulas (15.8*a*) and (15.8*b*) to take a glance at their origin. The idea of regression came first and the correlation method followed. It began with Sir Francis Galton, who was making some studies of heredity suggested by implications of the theories of evolution put forth by his even more illustrious cousin, Charles Darwin.

When Galton studied the relation of heights of offspring to the heights of their parents, he began by preparing a scatter diagram; perhaps the first. In order to put parents and their children on a common measuring scale, he converted all heights to standard scores. As the reader already knows, this meant expressing each person's height as a ratio of his deviation from his group mean to the standard deviation of that group dispersion. The unit for the offspring's scale and also for the parents' scale was then one σ. Figure 15.5 shows the type of figure Galton drew.

Galton next computed the means of offspring's heights (in z scores) corresponding to certain fixed parents' heights (in z scores). As we saw in the example earlier in this chapter when the same operations were performed (but with raw scores), he found that the means of columns fell along a straight-line trend. To him, incidentally, one striking phenomenon was that the means of offspring's heights did not increase as rapidly as did the parents' heights. Each mean height of offspring deviated less from their general mean than the height of the parents from which they came deviated from their mean. This "falling back" of heights of offspring toward the general mean has been called the *law of filial regression.* It is merely the phenomenon of imperfect correlation. Had the correlation between children and parents in height been perfect, the regression would have been as shown by the dotted line in Fig. 15.5. The correlation was actually about +.50, and the obtained regression line was as shown.

Origin of the Coefficient of Correlation.—Galton wanted a single value which would express the amount of this regression phenomenon in any particular relationship problem. Karl Pearson solved the problem in terms of the formula to which his name is attached. The steps were somewhat as follows. Galton's own idea was to use the slope of the regression line as the index of relationship, because the steeper the slope, the closer the agreement between two variables. The slope of the regression line in Fig. 15.5, as in any coordinate plot, is the ratio of the increase in Y corresponding to a certain increase in X. From the plot we see

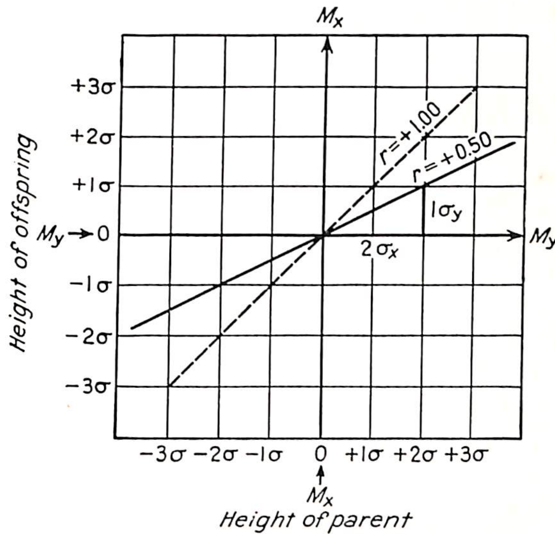


FIG. 15.5.—Diagram showing the relation of the Pearson product-moment coefficient of correlation to the slope of the regression line when scores in both X and Y are in standard-score units.

that as X changes 2σ (from the mean to $+2\sigma$, as shown), Y changes only 1σ. The slope is ½ or .5. This was Galton's coefficient of regression, which received the symbol r for that reason. That symbol has remained. The Pearson r *is* the slope of the regression line when both Y and X are measured in standard-deviation units. In this case, it can be shown that

$$r_{yx} = \frac{\Sigma z_y z_x}{N} \quad \text{(Pearson } r \text{ from standard measures)} \qquad (15.10)$$

In other words, r is an average of all the cross products of standard measures.

Derivation of the Regression Equations.—Since r is the slope of the regression line when standard measures are used, the equation for this

situation is

$$z_{y'} = r_{yx}z_x \qquad \text{(Regression equation with standard measures)} \qquad (15.11)$$

Here we use $z_{y'}$ with the prime to denote a predicted value as distinguished from the actual value. From this beginning, let us work toward the regression equations in raw-score form (formulas 15.8*a* and 15.8*b*). The next step is to express these standard measures as deviations, y' and x. Since $z_x = x/\sigma_x$ and $z_{y'} = y'/\sigma_y$ (σ_y is the unit of the $z_{y'}$ values as well as of the z_y values), the equation becomes

$$\frac{y'}{\sigma_y} = r_{yx}\frac{x}{\sigma_x} \qquad (15.12)$$

If we multiply this equation through by σ_y, we have

$$y' = r_{yx}\left(\frac{\sigma_y}{\sigma_x}\right)x, \qquad \text{(Regression equation with deviation scores)} \qquad (15.13a)$$

or

$$y' = b_{yx}x \qquad (15.13b)$$

Equation (15.13*b*) shows that the same *b* coefficient applies to deviation scores as that applying to raw scores (see formula 15.8*a*). It also shows that since the means of x and y are zero, the regression lines will pass through both of them without having an *a* coefficient in the equation.

One more step is needed to arrive at the raw-score type of regression equations. Going back to equation (15.12), if we next convert x to its equivalent, $X - M_x$, and y' to its equivalent, $Y' - M_y$ (M_y is the mean of the Y' values as well as of the y values), we have

$$\frac{Y' - M_y}{\sigma_y} = r_{yx}\left(\frac{X - M_x}{\sigma_x}\right) \qquad (15.14)$$

Multiplying through by σ_y, we have

$$Y' - M_y = r_{yx}\left(\frac{\sigma_y}{\sigma_x}\right)(X - M_x).$$

And transposing M_y,

$$Y' = r_{yx}\left(\frac{\sigma_y}{\sigma_x}\right)(X - M_x) + M_y,$$

which is identical with formula (15.8*a*).

Regression Coefficients from Ungrouped Data.—When data have not been grouped in class intervals, the derivation of the *b* coefficient requires another formula, which reads

$$b_{yx} = \frac{N\Sigma XY - (\Sigma X)(\Sigma Y)}{N\Sigma X^2 - (\Sigma X)^2} \quad \text{(Regression coefficient directly from data)} \qquad (15.15)$$

When this formula is applied to the data in Table 8.3, we have

$$b_{yx} = \frac{4{,}720 - 4{,}550}{6{,}240 - 4{,}900} = \frac{170}{1{,}340} = .127$$

The a coefficient is obtained by means of formula (15.7a) and is solved as follows:

$$a_{yx} = 6.5 - (7.0)(.127) = 6.5 - .89 = 5.61$$

The regression equation is therefore $Y' = 5.61 + .127X$. The equation for the regression of X on Y can be obtained by similar operations, substituting Y for X, and vice versa, in formula (15.15). The solution for the illustrative problem is

$$b_{xy} = \frac{4{,}720 - 4{,}550}{5{,}330 - 4{,}225} = \frac{170}{1{,}105} = .154$$

and

$$a_{xy} = 7.0 - (6.5)(.154) = 7.0 - 1.0 = 6.0$$

Checking the b coefficients, $b_{yx}b_{xy} = (.127)(.154) = .0196 = r^2$, which is in agreement with r^2 as previously known (see Table 8.3).

Predictions from Regression Equations.—As an illustration of how a regression equation is applied in prediction, let us assume some values of X and find the corresponding Y' values. Because in the preceding methods of prediction we predicted Y's corresponding to midpoints of the intervals of X, let us do the same here for the sake of comparison, remembering that we might have chosen any values of X that we pleased. Table 15.4 gives the X values and their corresponding Y' values. When X is 62, Y' is 105.1, and when $X = 97$, $Y' = 127.2$, etc. It is interesting to compare these particular predictions with the means of the columns, which are given in the third row of Table 15.4. The discrepancies will be

TABLE 15.4.—PREDICTIONS OF Y FROM X AND X FROM Y BY MEANS OF REGRESSION EQUATIONS*

$Y' = 0.630X + 66.05$

If X =	62	67	72	77	82	87	92	97
Y' =	105.1	108.3	111.4	114.6	117.7	120.9	124.0	127.2
M_c =	107.0	105.5	114.9	114.5	116.4	120.3	124.0	132.0

$X' = 0.591Y + 10.02$

If Y =	97	102	107	112	117	122	127	132	137
X' =	67.3	70.3	73.3	76.2	79.2	82.1	85.1	88.0	91.0
M_{row} =	67.0	70.3	74.0	75.9	78.6	83.2	85.8	83.7	97.0

* The data involved are from the two examinations correlated in Table 8.5. The means of the columns and rows are obtained from Table 15.2.

found very small as a rule. Granting that the column means are generally not very reliable because of small samples, we may feel more assurance in the Y' predictions because they are determined from the trend of the entire data rather than by small samples in separate columns. The predictions of X' from Y are given in the second section of Table 15.4 and are compared with the means of the rows as a matter of interest.

As a practical means of prediction, a graphic method will often be the most suitable procedure. If the regression lines are drawn as in Fig. 15.4 on cross-section paper, for any value of X on the base line, one can follow vertically up to the regression line and note the corresponding Y value at this point. One can read to the nearest unit with sufficient accuracy for practical work. The drawing of the regression line is simple in that two points determine the position of a line. One point can be at the two means, which will serve for both regressions. Another point for the regression of Y on X might be at $X = 60$, $Y = 103.85$; and a third point, for checking purposes, might be at $X = 100$ and $Y = 129.05$. For the regression of X on Y, points might be located conveniently at $Y = 100$, $X = 69.12$, and $Y = 130$, $X = 86.85$.

Standard Errors of the Estimates.—We previously saw (see Table 15.3) that the errors of prediction ($Y - Y'$ in the one case and $X - X'$ in the other) can be squared, summed, averaged and then the square root extracted in order to obtain the standard error of the discrepancies between observed values and predicted values. There we computed the standard error of the estimate from the discrepancies themselves; here we shall see that it is not necessary to compute the errors of prediction.

When we have predicted on the basis of regression equations, we can estimate the margin of error of prediction, as given by σ_{yx} (or by σ_{xy})

from the coefficient of correlation. The formulas are

$$\sigma_{yx} = \sigma_y \sqrt{1 - r^2_{yx}} \qquad (15.16a)$$

and (Standard error of estimate computed from r)

$$\sigma_{xy} = \sigma_x \sqrt{1 - r^2_{xy}} \qquad (15.16b)$$

in both of which the terms are now well known. It will be seen that the two equations are the same except for the use of σ_y when we are predicting Y and of σ_x when we are predicting X (for $r_{yx} = r_{xy}$). The two standard deviations are multiplied by the common factor $\sqrt{1 - r^2}$. This factor is always less than 1.00 and *gives us an estimate of the reduction in errors of prediction from knowledge of correlated measurements as compared to errors of prediction without that knowledge.* When r is zero, this element equals 1.00, and then $\sigma_{yx} = \sigma_y$, and $\sigma_{xy} = \sigma_x$. In other words, when $r = 0$, there is no basis for prediction. When $r = 1.0$ (or -1.0) the element reduces to zero, and so does the standard error of estimate. This coincides with the expectation that the margin of error of prediction is zero when the correlation is perfect.

Interpretation of an Obtained Standard Error of Estimate.—The interpretation of the standard error of the estimate when r is neither zero nor 1.00 is somewhat as follows. Like any standard deviation, σ_{yx} can be referred to the normal curve of distribution. For the examination problem,

$$\sigma_{yx} = 7.85 \sqrt{1 - .3721} = 6.22$$

and

$$\sigma_{xy} = 7.60 \sqrt{1 - .3721} = 6.02$$

No matter in what part of the measuring scale we are predicting (within the range of obtained scores, naturally) we assume that the margin of error is the same. When we predict Y from X, the average dispersion of observed measurements about Y' is given by a σ of 6.22. We expect two-thirds of the observed cases to lie within the limits of plus or minus 6.22 from Y'. This situation is illustrated graphically in Fig. 15.6. There we have the regression line, along which the predicted Y's lie, and in dotted lines we have the limits of one σ_{yx} on either side of it. Had we plotted a point for every individual, we should have expected about two-thirds of them to fall between the two dotted lines. To make a particular prediction, when $X = 90$, $Y = 122.8$. The odds are 2 to 1 that any individual whose X score is 90 will not fall below 116.6 or go above 129.0. We could state other odds for a divergence of 2σ either way or any other distance. It all depends upon our purposes.

We could prepare a similar diagram showing the limits of the middle two-thirds of the individuals about the regression of X on Y, and we could interpret the errors of prediction in a similar manner. It will be noted that the margin of error as given by σ_{xy} is 6.02, or 0.2 smaller in predicting in the other direction, *i.e.*, X from Y, but this is merely because σ_x is smaller than σ_y. The *percentage of error is the same in the two cases.* The ratio of σ_{yx} to σ_y is exactly the same as the ratio of σ_{xy} to σ_x, and that ratio is given by the factor $\sqrt{1 - r^2}$. This factor we will meet again with a name attached to it (see formula 15.21).

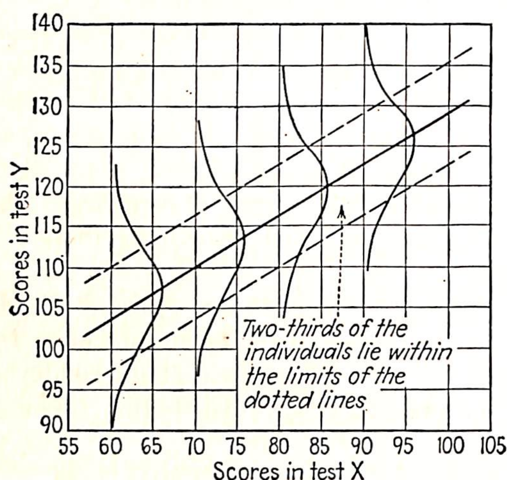


FIG. 15.6.—The line of regression of Y upon X, showing the range of observed values expected in Y in separate categories of score values on X. Parallel dotted lines above and below the regression line at a vertical distance of one standard error of the estimate each way mark off the region within which we expect two-thirds of the observed values to be.

The Regression Line as a Mean.—One way of looking at the regression line is to regard it as a moving average; a moving arithmetic mean. Like the arithmetic mean of any sample, the regression line satisfies the principle of least squares. The regression coefficients are so determined by the data that the sum of the squares of the deviations of observed points from the line is a minimum. Other lines might describe the trend of relationship nearly as well, but only the one line satisfies the principle of least squares. It is reasonable that if the line is a mean, the deviations from it should be measured by a standard deviation. That standard deviation is the standard error of estimate.[1]

[1] For an excellent critical discussion of regression effects in research problems see Thorndike, R. L. Regression fallacies in the matched group experiment. *Psychom.* 1942, **7,** 85–102.

Correction of a Standard Error of Estimate for Bias.—In smaller samples (N is less than 50) it would be well to make a correction in σ_{yx} (or σ_{xy}) before applying it to the population. The change can be made by the formula

$$_c\sigma_{yx} = \sigma_{yx}\sqrt{\frac{N}{N-2}} \quad \text{(Correcting } \sigma_{yx} \text{ for bias)} \quad (15.17)$$

where N is the number in the sample. The correcting can be done as well in the original computation, as follows:

$$_c\sigma_{yx} = \sigma_y\sqrt{(1 - r^2_{yx})\left(\frac{N}{N-2}\right)} \quad (15.18)$$

The Reliability of a Regression Coefficient.—The b coefficient in the regression equation has its sampling error, like all statistics. This is estimated by

$$\sigma_{b_{yx}} = \frac{\sigma_{yx}}{\sigma_x\sqrt{N}} \quad (15.19)$$

or by (Standard error of a regression coefficient)

$$\sigma_{b_{yx}} = \frac{\sigma_y}{\sigma_x}\sqrt{\frac{1-r^2}{N}} \quad (15.20)$$

The $\sigma_{b_{xy}}$ would be the same, except for changing the x and y subscripts around. For our examination problem

$$\sigma_{b_{yx}} = \frac{6.22}{(7.60)(9.3274)} = \frac{6.22}{70.88824} = .088$$

We may say that the odds are 2 to 1 that the obtained b_{yx} of .63 does not deviate from the population b_{yx} by more than .088. There is very little chance that the true b coefficient here is zero.

The Correlation Coefficient and Accuracy of Prediction

The chief index of goodness of prediction of measurements thus far in this discussion has been the standard error of estimate. It has been shown how the latter is closely related to the coefficient of correlation. As r increases, the standard error of estimate decreases. There are other ways in which r and some of its derivatives can be used to indicate accu-

racy of prediction. Three of the common derivatives are the *coefficient of alienation*, the *index of forecasting efficiency*, and the *coefficient of determination*. Each has its unique story to tell about the closeness of correlation between two things and about the utility of predictions.

The Coefficient of Alienation.—Whereas r indicates the strength of relationship, the *coefficient of alienation*, k, indicates the degree of *lack* of relationship. By formula,

$$k = \sqrt{1 - r^2} \quad \text{(Coefficient of alienation computed from } r\text{)} \quad (15.21)$$

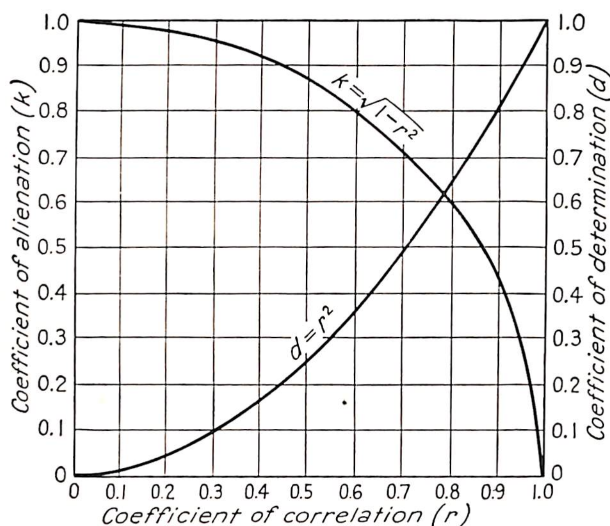


FIG. 15.7.—Chart showing k (coefficient of alienation) and d (coefficient of determination) as functions of r (coefficient of correlation).

Squaring both sides of this equation, we have

$$k^2 = 1 - r^2$$

And transposing, we have

$$k^2 + r^2 = 1.00$$

Thus, although we might have expected k plus r to equal 1.00, it is rather the sum of their squares that equals 1.00. If r is .50, k is *not* also .50 but .886. When r is .50, then, the degree of relationship is less than the degree of *lack* of relationship. It is when $r = .7071$ that relationship and lack of relationship are equal, for k also then equals .7071. Then $r^2 + k^2 = .50 + .50 = 1.00$. Other values of k for different sizes of r can be found in Table 15.5. Figure 15.7 shows pictorially the functional relationship between k and r. Students of mathematics will recognize

TABLE 15.5.—INDICATORS OF THE IMPORTANCE OF COEFFICIENTS OF CORRELATION

r_{xy}	k_{xy} Coefficient of alienation	$100(1 - k_{xy})$ Percentage reduction in errors of prediction of Y from X	$100r^2_{xy}$ Percentage of variance accounted for
.00	1.000	0.0	0.00
.05	.999	.1	0.00
.10	.995	.5	1.00
.15	.989	1.1	2.25
.20	.980	2.0	4.00
.25	.968	3.2	6.25
.30	.954	4.6	9.00
.35	.937	6.3	12.25
.40	.917	8.3	16.00
.45	.893	10.7	20.25
.50	.866	13.4	25.00
.55	.835	16.5	30.25
.60	.800	20.0	36.00
.65	.760	24.0	42.25
.70	.714	28.6	49.00
.75	.661	33.9	56.25
.80	.600	40.0	64.00
.85	.527	47.3	72.25
.90	.436	56.4	81.00
.95	.312	68.8	90.25
.98	.199	80.1	96.00
.99	.141	85.9	98.00
.995	.100	90.0	99.00
.999	.045	95.5	99.80

the relationship $r^2 + k^2 = 1.00$ as the equation for a circle with a radius of 1.00. The diagram shows only positive values of r and k.[1]

Sometimes we wish to stress the point of independence between two things rather than their closeness of agreement. In such instances, we present k as well as r. Besides being related to r, k is also related to other indices of goodness of prediction to be mentioned next.

[1] The relation of k to r is the same as that of the sine of an angle to the cosine of that angle. Values of k corresponding to known values of r can be found by using Table J in Appendix B.

The Index of Forecasting Efficiency.—In the formula for the *SE* of the estimate, $\sigma_{yx} = \sigma_y \sqrt{1 - r^2_{yx}}$, we can now see that the factor under the radical, $\sqrt{1 - r^2_{yx}}$, is really the coefficient of alienation. We could rewrite the formula as $\sigma_{yx} = \sigma_y k_{yx}$. If we were to multiply k by 100, we should have the percentage σ_{yx} is of σ_y. When $r = .61$, as in our recent illustration, $k = .7924$. The *SE* of the estimate in this problem is 79.24 per cent of the observed dispersion of observations. Our margin of error in predicting Y *with* knowledge of X scores is about 79 per cent as great as the margin of error we would make *without* knowledge of X scores. For then we predict every Y to be the mean of the Y's, and the *SE* of the prediction then equals σ_y. The *reduction* of our margin of error is 100 minus 79.24, or 20.76 per cent. The *index of forecasting efficiency* is defined as the percentage reduction in errors of prediction by reason of correlation between two variables. The general, simplified formula is

$$E = 100(1 - \sqrt{1 - r^2}) \qquad \text{(Index of forecasting efficiency)} \qquad (15.22)$$

or

$$E = 100(1 - k)$$

The calculation of E is facilitated by Table 15.5, where many of the E values are given for corresponding r's. Inspection will show that r must be as high as about .45 before E is 10 per cent. When a test has a validity coefficient of .45, the size of errors of prediction, on the whole, is only 10 per cent less than that we would have without knowledge of test scores but with knowledge of the mean criterion measure. Taken at its face value, this does not seem like much of a gain. There are situations, however, in which, as will be shown later, a gain of even less might be of practical importance.

Better tests, with validity coefficients of .60, have an E of 20 per cent, and still better tests, when $r = .75$, have an E of about 34 per cent. Although these efficiencies may also seem small, we must treat them in a relative, not an absolute sense. It is probable that the efficiency of predictions based upon the average unsystematic interview is less than 5 per cent. With this as our base, the picture of efficiency of tests looks much better.

Figure 15.8 shows graphically the functional relationship between E and r. The range of r's from .3 to .8 is marked off as representing the level of validity coefficients usually found for useful predictive instruments in psychological and educational practices. Tests rarely show correlations greater than .8 with practical criteria, and those correlating less than .3 are usually of limited value when used alone. In a battery

to which they make a unique contribution it may still be worth while to use them. The corresponding limits on the scale of E are 4.6 and 40.

The Coefficient of Determination.—Another mode of interpretation of r is in terms of r^2, which is called the *coefficient of determination*. This statistic is also sometimes symbolized as d. The coefficient r^2 gives us (when multiplied by 100) the percentage of the *variance* (see Ch. 5) in Y that is associated with or determined by variance in X. When $r = .50$, the percentage of the variance in Y that is accounted for by variance in X is 25, or one-fourth. To account for half the variance of any set of measurements, the r with another variable would have to be .7071. The

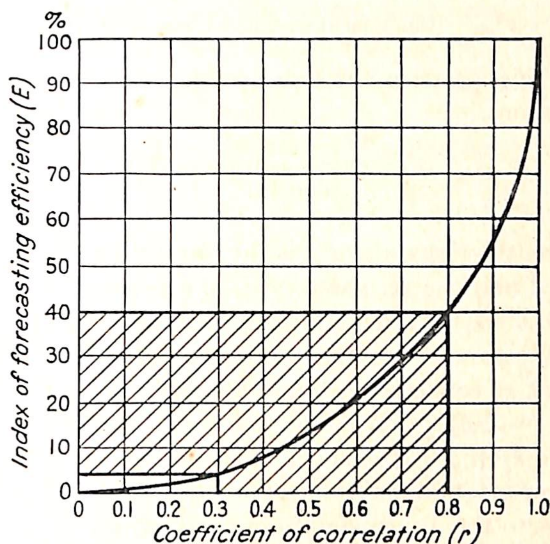


FIG. 15.8.—E (index of forecasting efficiency) as a function of r.

proportion of the variance in Y *not* determined by or associated with variance in X is given by k^2, which is called the *coefficient of non-determination*. These statements about determination of Y by X are reversible and apply equally well to determination of X by Y. We should speak of *determination* of one thing by another, however, only when a causal relationship can be logically defended; otherwise the expression *associated with* or *accounted for* (by way of prediction) is better. In Table 15.5, several of the $100r^2$ values are given for corresponding r's. In Fig. 15.7 is presented graphically the functional relationship between d and r.

Predicted and Nonpredicted Variances.—The coefficient of determination, as well as its relations to r, k, and other statistics, can best be clarified by introducing another new idea. The total amount of variance in the pre-

dicted variable, Y, we denote by σ^2_y. We can think of this variance as being broken down into two independent components, the predicted and the nonpredicted portions. The predictions of Y, which we have called Y', have their dispersion and their variance which are denoted by $\sigma_{y'}$ and $\sigma^2_{y'}$, respectively. The standard deviation $\sigma_{y'}$ would be computed from the deviations of the predicted values about the mean of the Y values, M_y. The amount of nonpredicted variance is indicated by the square of the standard error of estimate (σ^2_{yx}). This statistic is computed from the deviations of the obtained Y values from the regression line (or from the predicted Y values). The two component variances of σ^2_y are therefore

$$\sigma^2_y = \sigma^2_{y'} + \sigma^2_{yx} \quad \text{(Component variances in the predicted variable)} \quad (15.23)$$

If we divide this equation through by σ^2_y, we will have everything in terms of proportions.

$$\frac{\sigma^2_y}{\sigma^2_y} = \frac{\sigma^2_{y'}}{\sigma^2_y} + \frac{\sigma^2_{yx}}{\sigma^2_y} = 1.0 \quad \text{(Total variance as the sum of two proportions)} \quad (15.24)$$

The first term on the right, $\sigma^2_{y'}/\sigma^2_y$, is the proportion of the variance in Y that is predicted and the second term is the proportion of the variance that is not predicted. We have already defined r^2 as the proportion of predicted variance and k^2 as the proportion of nonpredicted variance. This means that r^2 equals $\sigma^2_{y'}/\sigma^2_y$ and that k^2 equals σ^2_{yx}/σ^2_y, and that $r = \sigma_{y'}/\sigma_y$ and $k = \sigma_{yx}/\sigma_y$. We therefore have some new concepts of r and k. We can say that r is the ratio of the dispersion of predicted values to the dispersion of obtained values and that k is the ratio of the dispersion of errors to the dispersion of obtained values.

Effectiveness of Selection Tests

Although the coefficient of correlation and its derivatives, k, E, r^2, and σ_{yx}, are all accurate and meaningful ways of interpreting the goodness of predictions, and they serve well for those who know how to use them, in some practical situations they leave something to be desired. To quote them to the layman may earn the investigator a cool reception and an empty stare. Even the statistically informed test expert may find it desirable at times to cast his conclusions in other terms. This is true, particularly, when we are dealing with selection tests.

Those concerned with the administrative problems of selecting personnel by means of tests find that a different kind of enlightenment is desirable than that provided by the statistics in question. It is one thing to know that by the use of this test score, or this composite score, errors of pre-

diction are reduced 15 per cent. But what does this mean with regard to the number of applicants one must examine, and what proportion one must accept for training in order to have a certain number of successful employees at the end of training? With the same number of applicants selected, how many more satisfactory ones will we have with the aid of the selection test than we would have had without it? Even if we could get the employer to grasp the idea of the index of forecasting efficiency as an abstract indicator of amount of gain achieved by the test, to most laymen the E values actually reached by most test procedures sound very unimpressive because laymen generally lack the proper experience to evaluate them. For these reasons, several suggestions have been made in recent years for more realistic and fruitful ways of evaluating selection tests. One of these will be described in some detail and the others mentioned in principle.

Determiners of Effective Selection.—Everything else being equal, validity coefficients (and statistics derived from them) are accurate indices of the effectiveness of selection tests. It has been pointed out, however, that the correlation of a test with a practical criterion is not the only thing to be considered when practical decisions must be made. The practical utility of tests in any training or job situation depends upon other factors than the validity of the test or test battery. It depends upon the percentage of employees who would have succeeded if testing had not been applied in selection. It also depends upon the percentage of the applicants who are selected by means of the tests.

The Taylor-Russell Method.—Taylor and Russell have rationalized the problem in a clear manner.[1] Following their exposition of the matter, the selection situation with tests is described in Fig. 15.9. The X axis represents the scale of test scores and the vertical axis represents the scale of the training or job criterion. Let us assume that the correlation between test and criterion is about .50. The ellipse describes the dispersion of individuals in this two-dimensional surface. On the X scale a point X_c is marked. This is an arbitrary critical or qualifying score on the test. Individuals with scores above X_c are selected and those with scores below X_c are rejected.

Without selection on the basis of the test, a certain percentage of the accepted applicants would have succeeded. We assume a continuous variable for the criterion as well as for the test. The point Y_c is an arbitrary critical criterion value above which the verdict is success and below which the verdict is failure. By drawing lines at X_c and Y_c parallel

[1] Taylor, H. C., and Russell, J. T. The relationship of validity coefficients to the practical effectiveness of tests in selection. *J. appl. Psychol.*, 1939, **23**, 565–578.

to axes Y and X, respectively, we divide the population into four kinds of individuals defined as follows:[1]

A—Individuals who if selected would succeed.

B—Individuals who would be rejected but who if allowed to compete would succeed.

C—Individuals who if selected would fail.

D—Individuals who would be rejected and who if allowed to compete would fail.

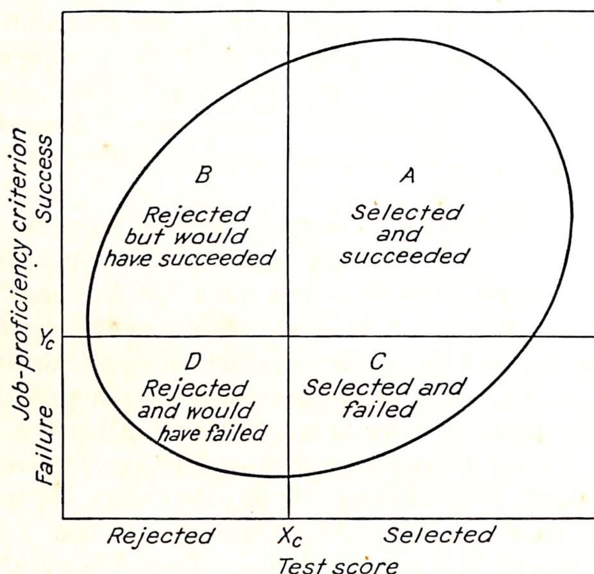


FIG. 15.9.—Correlation surface subdivided by a critical score (X_c) which separates the population into selected and rejected groups of individuals on the basis of test results, and by a critical criterion value, (Y_c) which separates the same population into successful and unsuccessful in a job assignment.

Success Ratios and Selection Ratio.—It is clear that the *A* and *D* people are correctly predicted under these conditions and the *B* and *C* people are incorrectly predicted. We have thus reduced the prediction problem to one of prediction of (quantitative) attributes from (quantitative) attributes. The evaluation of predictions in this form could be carried out much as was described earlier. Here, however, the problem is much more complicated, because we have to consider different division points on the success scale as well as different critical scores for selection on the test scale. In attribute-prediction problems the division points are usually fixed by the nature of things.

[1] The letter symbols—*A*, *B*, *C*, *D*—are defined somewhat differently than by Taylor and Russell. Here they have been made more consistent with the corresponding categories—*a*, *b*, *c*, *d*—in the usual 2 × 2 contingency table.

We are now ready to consider two new concepts proposed by Taylor and Russell. One is the *success ratio* and the other the *selection ratio.* The success ratio is the proportion of accepted candidates who would be successful. There would be a certain success ratio *without* the use of selection tests, and another success ratio *with* the use of tests, provided the tests have any validity at all, and provided *some* selection occurs. The selection ratio is the proportion of all applicants examined who are accepted. In terms of symbols and equations, the success ratio without the use of tests is

$$S_o = \frac{A + B}{A + B + C + D} = \frac{A + B}{N} \quad \text{(Success ratio without the use of selection tests)} \qquad (15.25)$$

where letters A, B, C, and D, are defined as in Fig. 15.9. When there has been selection on the basis of a valid test,

$$S_t = \frac{A}{A + C} \quad \text{(Success ratio with the use of tests)} \qquad (15.26)$$

The selection ratio is

$$p_s = \frac{A + C}{A + B + C + D} = \frac{A + C}{N} \quad \text{(Selection ratio)} \qquad (15.27)$$

Favorable Success Ratios (before Selection).—A few examples will illustrate the fact that effectiveness of selection by tests depends upon the success ratio that would prevail without that selection. It is obvious that if all trainees or employees would be satisfactory without the use of selection tests, there would be little excuse for using them. The chances of improving matters by this approach would be nil, except as the *quality* or average production of satisfactory personnel were raised as a result. When the success ratio without tests is very low, there is much room for improvement, and with valid tests some improvement is bound to occur.

Consider Fig. 15.10 in this connection. There four special situations are shown; cases of high and low test validity combined with high and low success ratios. In diagram I, the success ratio is high. One could move the critical score over a considerable range without changing the success ratio very much, until the selection ratio became very small. In diagram II, even where the correlation is high, a change in the cut-off score would disqualify very few potential failures, and eliminating even a few would result in losing many more potentially successful candidates. In diagrams III and IV, the success ratios are very small. In diagram III, even a small number of rejections would disqualify many potential

failures with little or no loss of potential successes. This is even more true where the validity of the test is much higher, as shown in diagram IV. In general, then, *we stand to gain most when success ratios without testing are small.*

Favorable Selection Ratios.—If the number of applicants relative to the number of places to fill is small there is, of course, not much opportunity for selection. In the limiting case, if no one could be rejected there

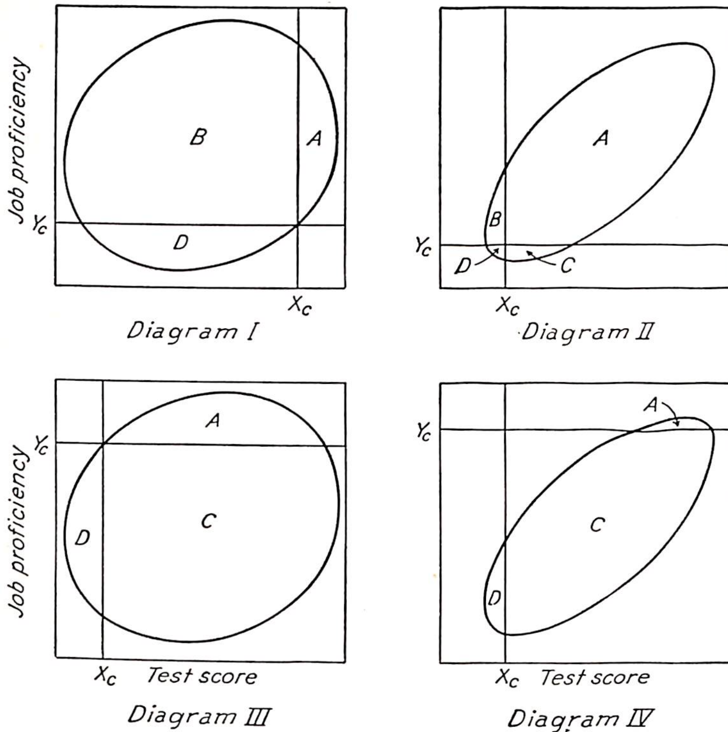


FIG. 15.10.—Diagrams similar to that in Fig. 15.9, with different combinations of selection ratio and success ratio (for definition of these ratios, see the text), and different degrees of validity.

would be no use of selection tests. On the other hand, if there are many more applicants than places, and if one can then skim the "cream" off the top of the applying group, the chances of improving the quality of accepted personnel would seem to be great. This presupposes a method whereby the "cream" can be properly recognized. A valid test does that. But how valid must a test be before there is sufficient recognition of top talent?

Figure 15.10, diagram III, shows that even a test of rather low validity may be effective in skimming the "cream" in a negative way. That is,

it can do much to eliminate failures. One could move the critical score a considerable distance and still reject several times as many potential failures as he would lose among potential successes. From diagram I, however, we get the suggestion of a warning that if the qualifying score is set too high in a test of low validity we may be losing some of the very best qualified. We cannot press refined decisions of this kind too far on the basis of these diagrams because the populations are not uniformly distributed throughout the elliptical areas; they thin out around the margins. The general tendencies, however, should be clear.

It is clear from what has been said above that a test of low validity may be very useful in selection under a favorable set of conditions. Those conditions include certain combinations of success ratios and selection ratios. It can also be seen that even a test of high validity may be of little or no value if the conditions are unfavorable. Consider diagram II, in which the success ratio is very high. One could not eliminate many potential failures without losing many more satisfactory personnel. The higher the critical score, however, the more satisfactory the successful personnel would tend to be. It depends upon whether we are interested in *numbers* of successful individuals or in average quality. There are administrative questions of balance, also. It might be disadvantageous to take on at one time a whole class of prima donnas!

Some numerical examples may be given to illustrate the points just made concerning favorable success and selection ratios. Let us assume a validity coefficient of .60, a typical value for good selection batteries. Let us also assume normal distributions in both test and criterion. If the success ratio S_o is .95, by rejecting 40 per cent of the applicants we could achieve a success ratio S_t of .99. This is an improvement of only about 4 per cent over the results without the tests. Compare this with the index of forecasting efficiency which is 20 per cent when $r = .60$. To bring the S_t up to 1.00, approximately, we would need to reject at least 60 per cent of the applicants. In either case, we reject about 10 applicants to gain one more successful individual. Rejections beyond 60 per cent would gain us practically nothing.

Let the success ratio S_o be .05, and what is the result? A rejection of 55 per cent of the applicants would net an increase of .05 in the success ratio, a gain of 100 per cent. By rejecting as many as 95 per cent the S_t could be raised to .30. This is a gain of 500 per cent. Compare these percentage gains with the index of forecasting efficiency of 20.

To take less extreme instances of S_o, let us assume ratios of .80 and .20, with r still equal to .60. With the high S_o of .80, we need to reject about 60 per cent in order to raise S_t to .95, a gain of 17.5 per cent. With the

low S_o of .20, the rejection of 60 per cent yields a success ratio of .38, a gain of 90 per cent.

A Graphic Chart of Relations of S_t to Selection Ratio.—Figure 15.11 shows, for the situation when the validity coefficient is .60, the change in success ratio S_t as the selection ratio changes. Each curve represents a different initial or basic success ratio, S_o. Taylor and Russell provide tables which record these same relationships for various validity coefficients and Guilford and Michael provide charts similar to Fig. 15.11 for other validity levels.[1]

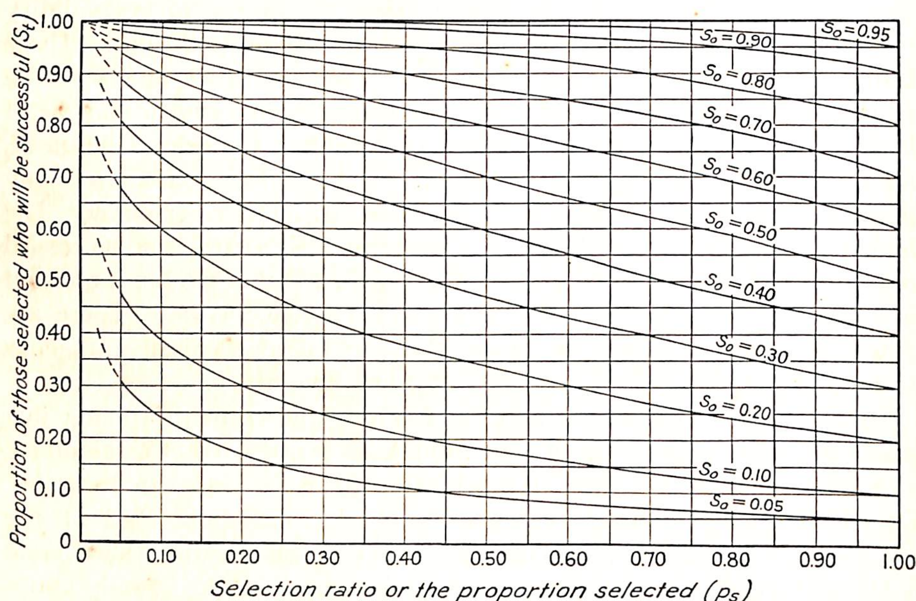


FIG. 15.11.—Chart relating success ratio to selection ratio when the validity coefficient is .60.

Indices-of-Improvement Methods.—In the Taylor-Russell method of test evaluation our attention is concentrated upon *numbers* and *percentages* of successful individuals. We ask what is the percentage increase in the numbers of satisfactory personnel, without specifying anything about the *degree* of satisfaction. Much depends upon the placing of a passing point on the criterion scale and an ignoring of the fact that success is a graded variable. In terms of planning in selection and training programs, particularly in military situations, where numbers of recruits may be liberal and standards of passable satisfaction are readily established, this kind

[1] Taylor and Russell, *op. cit.;* Guilford, J. P. and Michael, W. B. Prediction of categories from measurements. Beverly Hills, Calif.: Sheridan Supply Co., 1949.

of evaluation of a selection instrument or program is adequate and well adapted. There are other procedures, however, that concentrate more upon the fact of graded excellence in criterion measures, and which involve thinking in terms of work output of personnel. The worth of a selection program is established if we can demonstrate a certain percentage increase in production of some kind. If the criterion is measured in terms of absolute amounts of production of workers, we may ask, "What percentage improvement in production does test selection bring about?" The answer can then be balanced against the cost of the testing program.

The Jarrett Method.—Although the first suggestion for this kind of index of test evaluation was made by Richardson,[1] a more useful procedure was developed by Jarrett.[2] With somewhat different symbols than those used by Jarrett, his index of improvement can be computed by the formula

$$I = r_{yx}v_y\left(\frac{M_p - M_x}{\sigma_x}\right) \quad \text{(Percentage improvement in output for a selected group)} \qquad (15.28)$$

where r_{yx} = validity coefficient for the test.
v_y = index of variability of criterion scores given by the equation[3]

$$v_y = \frac{\sigma_y}{M_y} \quad \text{(Relative variability of measurements)} \qquad (15.29)$$

where M_p = mean of test scores for the selected personnel.
M_x = mean of test scores for all applicants.
M_y = mean of criterion measurements.
σ_y = standard deviation of the criterion measures.

If we may assume that the criterion measures are normally distributed, the last term in formula (15.28) is equivalent to the ratio y/p_s and we have

$$I = r_{yx}v_y\frac{y}{p_s} \quad \text{(Percentage improvement in output for a selected group in a normally distributed criterion)} \qquad (15.30)$$

[1] Richardson, M. W. The interpretation of a test validity coefficient in terms of increased efficiency of a selected group of personnel, *Psychom.*, 1944, **9**, 245–248.

[2] Jarrett, R. F. Percent increase in output of selected personnel as an index of test efficiency, *J. appl. Psychol.*, 1948, **32**, 135–145.

[3] The statistic v_y will be recognized as 1/100 of the coefficient of variation given in Ch. 5. Here, as well as there, the measurements must be in terms of a scale with an absolute zero point. Piecework scores, dollar values of output, and the like, qualify for the use of this statistic. Ratings would not qualify.

where p_s = proportion of applicants selected.
y = ordinate in unit normal distribution curve at point marking off p proportion of cases.

An inspection of formula (15.30) leads to some interesting inferences which agree with things already pointed out. With v_y and y/p constant, I is entirely dependent upon the validity of the test and directly proportional to it. With r_{yx} constant, I increases as v_y increases. That is, the more variable the criterion measures with respect to their mean, the greater is the improvement resulting from selection. It is reasonable that if all workers performed equally well there would be little use to attempt to discriminate among them by means of tests. The better they can be discriminated in terms of individual output, the better the chance there is of differentiating among them by means of predictive instruments. The factor y/p, as will be seen in Table G, is larger as p approaches .00 and smaller as p approaches 1.0. When $p = .01$ this ratio is about 100 times as large as when $p = .99$. This principle agrees with the one applying to the Taylor-Russell method: that the lower the selection ratio, the greater the benefit from selection.

Evaluation in Terms of Cost and Utility.—Berkson has recently developed a procedure which emphasizes a comparison of the *utility* of a test with its *cost*. Utility is defined as the percentage of potential failures that would be eliminated by the test. Cost is the percentage of potential graduates the test would eliminate. These definitions can be referred to Fig. 15.9. Utility would equal $100D/(C + D)$. Cost would equal $100B/(A + B)$. The indices are, of course, related to the positions of the cut-off score and to the success ratio. In comparing tests, Berkson uses a single index number based upon the average cost for all utilities. For details the reader is referred to Berkson's description.[1]

Selection When Regressions Are Nonlinear.—Previous discussions of selection by means of tests have assumed linear regression; the assumption that throughout the range, the higher the score the greater the average criterion performance of the individual. We should not leave the subject of selection without considering the case of curved regressions. Figure 15.12 shows in general form a type of regression that may be more common than has been realized.

There has been a common conclusion in the industrial-psychology literature that individuals of high intelligence are likely to do less well at highly routinized, repetitive tasks than individuals of lower intelligence. The effect may be due to lack of interest and to boredom on the part of

[1] Berkson, J. Cost-utility as a measure of the efficiency of a test, *J. Amer. stat. Assoc.*, 1947, **42**, 246–255.

the highly intelligent person, but for predictive purposes we do not particularly need to know the reasons. The fact of curved regression is undeniable and should be recognized in selection. It is likely that when the whole range of intelligence is studied in relation to job proficiency of many kinds, there will be found an optimal intelligence level for each kind of job. Curved regressions are often overlooked because the investigator fails to plot scatter diagrams, or because he has a restricted range in his population. In application for jobs, there is often enough self-selection

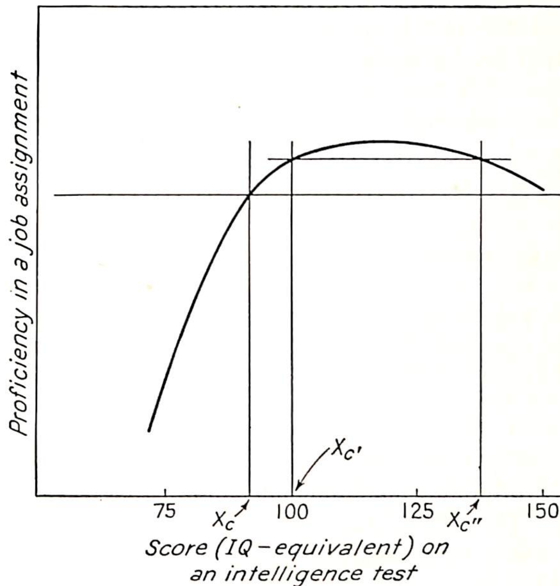


Fig. 15.12.—A curved regression of a job-proficiency criterion variable on the test-score variable X, showing that a high cutoff score may be needed in addition to a low one.

beforehand so that a limited range appears for examination. The resulting regression is therefore often linear within that range and some correlations are zero because in that range there is no upward trend in Y as X increases. In relating certain temperament-test scores to rated proficiency of administrators, for example, the writer has found a few undeniable signs of curvature, with the optimal score not at the top. Relations of other temperament scores to job proficiency measures in such routine tasks as cigar wrapping and stocking pairing reveal optimal scores below the average, that is, toward the extreme ordinarily denoted as poor personality traits.

Wherever curvature such as that shown in Fig. 15.12 is indicated by the data, two critical scores may be called for. If a cutoff score were

placed at X_c, then all the personnel above that point are apparently about equally good in terms of job proficiency. If the cutoff point were moved up to $X_{c'}$, however, there are individuals having scores at the upper end who are just as poor performers on the job as many below $X_{c'}$. A second critical point at $X_{c''}$ would eliminate the high-scoring but below-optimal performers. If selection were further restricted, it should be restricted from both directions.

The problems of evaluating selection devices when regressions are nonlinear are more complex than those we have already seen. None has been worked out for this kind of situation, but variations of methods already described would serve. The fundamental principles would be the same.

Exercises

1. Using the data of Table 14.10, predict the most probable score in the personality inventory for alcoholics and nonalcoholics, and for the two combined. What is the margin of error of prediction as made in these three ways?

2. Compute a standard error of estimate for the prediction problem in Exercise 1. What does it tell us?

3. What is the most probable total score for the passing and failing students represented in Table 13.4? What is the accuracy of prediction for each category? How much improvement from knowledge of category?

4. For Data 15*A*, find the best prediction of score in the Opposites test corresponding to each midpoint score in the Mixed-sentences test. Estimate the margin of error for each prediction and for the predictions taken as a whole.

5. Find the two regression equations for Data 15*A*. Make all possible checks as to internal consistency of your computations.

DATA 15*A*.—A SCATTER DIAGRAM FOR TWO MENTAL TESTS

Y (Opposites test in Army Alpha)	X (Mixed-sentences test in Army Alpha)								
	0–2	3–5	6–8	9–11	12–14	15–17	18–20	21–23	f_y
36–38								1	1
33–35							1	2	3
30–32				1	1	3	7	2	14
27–29						4	5	2	11
24–26			1	3	3	2	4	4	17
21–23			1		6	1	5	2	15
18–20		1	2	1	9	5	4		22
15–17	2	1	2	2	2	2	1		12
12–14	1	2	0	2	2	1			8
9–11	3	1	2	1	2				9
6– 8				1					1
f_x	6	5	8	11	25	18	27	13	113

6. Using the appropriate regression equation, make a prediction of score in the Opposites test corresponding to each midpoint score in the Mixed-sentences test. Compare these predictions with those obtained in Exercise 4.

7. Compute the two standard errors of estimate for Data 15*A*. What are the *amounts* of predicted and of nonpredicted variance in Y? What are the *proportions* of these two kinds of variances here?

8. Draw a diagram like Fig. 15.6 that applies to Data 15*A*. Draw another diagram like Fig. 15.4 showing the two regression lines.

9. Derive the statistics k, E, and r^2 for Data 15*A*. Interpret these findings.

10. Using formula (15.15), compute a regression equation for the first 10 pairs of scores for Parts V and VI in Data 8*A*.

CHAPTER 16

MULTIPLE PREDICTION

Multiple Correlation

Independent and Dependent Variables.—Thus far we have been dealing with correlations between two things at a time and the prediction of some variable Y from another variable X, or vice versa. Actual relationships between measured things in psychology and education are by no means so simple as that. One variable is found associated with, or dependent upon, more than one other variable at the same time. When we can think of some variables as being causes of another one, or even when we merely want to predict that one from our knowledge of several others that are correlated with it, we call the one variable the *dependent* variable and the ones upon which it depends the *independent* variables. The independent variables are so called because we can manipulate them at will or because they vary by the nature of things, and in consequence, we expect the dependent variable to vary accordingly.

Whether or not a certain color is liked depends upon several factors: its hue (whether yellow, red, or purple, etc.), its lightness (whether light, medium, or dark), and its chroma (saturation or density). The affective value of the color also depends upon its area, its use, and its background. We are here naming independent variables upon which the affective value of a color depends. In so far as each one is a determiner of agreeableness of color, it will exhibit some correlation individually with affective value. The size of any one of these correlations will depend upon the relative strength of that factor and also upon how well the other factors have been neutralized, as they should be in a good experimental situation.

A Graphic Picture of Multiple Dependence.—The idea of a dependence of one variable upon two others can be illustrated by Fig. 16.1. In that illustration is shown how the dependent variable, success in pilot training, is related both to aptitude scores and to chronological age. It requires a three-dimensional figure to show the relationships. The vertical dimension represents the dependent variable. Here it is measured in terms of percentage of graduates—not an ordinary way of measuring, but it will, nevertheless, show the principles involved. The two independent variables are shown as sides of the base, at right angles to each other. The

scale of chronological age is shown reversed for convenience, since the correlation between age and the training criterion was negative. Both independent variables are shown here in very coarse categories for the sake of a simpler diagram.

By noting rows of blocks (left to right) we can see how graduation rate changes with age for a relatively constant level of aptitude. By noting the columns of blocks (front to back) we can see how graduation rate changes with aptitude score for a relatively constant age level. The term

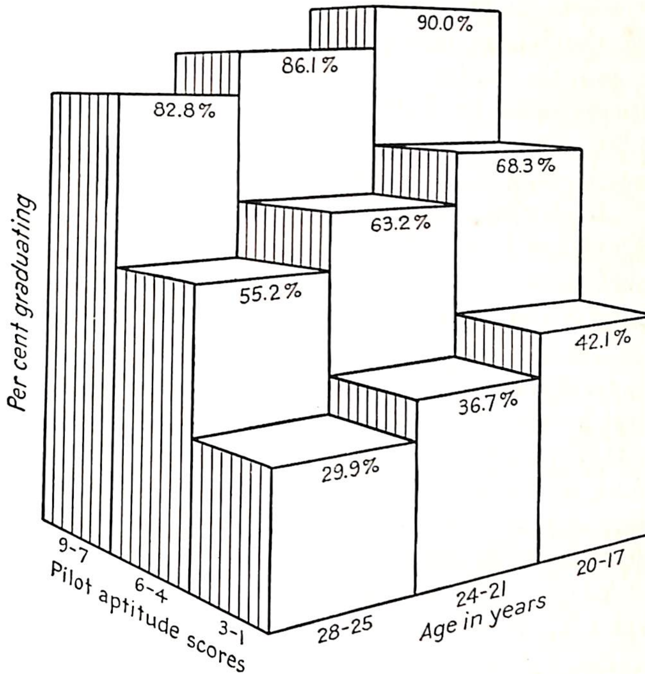


FIG. 16.1.—A multiple regression, with percentage graduating from pilot training as a function of both chronological age and aptitude score. (*Adapted from an unpublished report of Headquarters, AAF Training Command, Fort Worth, Texas.*)

constant covers an unusual range in this illustration, but with finer grouping on age and aptitude we would expect similar trends. It is obvious that the regressions for the criterion on aptitude are much steeper than those for the criterion on age. The difference would be even more apparent if we had the criterion in terms of a properly graded measurement scale. The correlation between aptitude scores and the criterion was much higher (approximately .55) than that between age and the criterion (approximately −.10). A very rough appreciation of the joint predictive value of aptitude score and age can be seen by noting the change of height

from the lowest block (29.9 per cent) to the highest (90.0 per cent). This change may be compared with those changes across columns alone or across rows alone. From this comparison we should expect better prediction from both independent variables than from either alone.

The Coefficient of Multiple Correlation.—When we are interested in the amount of correlation between a dependent variable and two or more others simultaneously, we are dealing with a multiple-correlation problem. The multiple coefficient of correlation indicates the strength of relationship between one variable and two or more others taken together. The multiple correlation is not merely the sum of the correlations of the dependent variable and the various independent variables taken separately. Obviously, there would be instances in which these would add up to more than 1.00. One reason is that independent variables themselves are usually overlapping (intercorrelated) and so duplicate one another to some extent. In this we see one important principle of multiple correlation. The multiple R is related to the intercorrelation of independent variables as well as to their correlation with the dependent variable. The interdependency of the determiners suggested for affective value of colors is probably not so apparent as in the case of factors related to achievement in college algebra. Here we can think of such predictive factors as intelligence-test scores and high-school marks, which being related duplicate one another to some extent in predicting achievement in college algebra. Hours of study and interest also bear much in common and so are not completely independent determiners of success in algebra.

A Multiple-correlation Problem.—In Table 16.1 are presented some data that call for the multiple-correlation solution. Four of the variables (X_2, X_3, X_4, and X_5) are all measures of things that supposedly determine academic success in college freshmen. X_1 is the dependent variable, or average freshman marks. It is customary to designate the dependent variable by X_1, though some authors, less often, call it X_0. An examination of Table 16.1 shows that the analogies test and high-school average mark have the highest correlation, when taken alone, with X_1, whereas the interest score X_5 has the lowest. The highest *intercorrelations* come between X_2, X_3, and X_4. All represent abilities of one kind or another, and their correlations with X_5 (interests) are generally lower. This gives promise that the interest scores will contribute something to the prediction of college marks that will not have been already contributed by the other variables, and so it should pay to include X_5 in the battery of predictive indices. As a matter of experience in psychological and educational predictions, it has been a common finding that it rarely pays to bring into a multiple-prediction situation more than four or five independent variables.

TABLE 16.1.—INTERCORRELATIONS AMONG FIVE VARIABLES, INCLUDING ONE INDEX OF SCHOLARSHIP AND FOUR PREDICTIVE INDICES (N = 174)*

Variable	X_2	X_3	X_4	X_5	X_1
X_2	—	.562	.401	.197	.465
X_3	.562	—	.396	.215	.583
X_4	.401	.396	—	.345	.546
X_5	.197	.215	.345	—	.365
X_1	.465	.583	.546	.365	—
M_x	19.7	49.5	61.1	29.7	73.8
σ_x	5.2	17.0	19.4	3.7	9.1

X_2 = arithmetic test in the Ohio State Psychological Examination, Form 10.
X_3 = analogies test in the same examination.
X_4 = an average grade in high-school work.
X_5 = student interest inquiry (measuring breadth of interest).
X_1 = an average grade for the first semester in university.

* These data were abstracted from the *Ohio State Coll. Bull.* 58, by L. D. Hartson, and have been used in this chapter by permission.

By the time that this many are combined, they have fairly well covered what any additional one can do for us. This is partly a consequence of the fact that good human qualities tend to go together (to be intercorrelated) and partly that our predictive indices tend to remain in the same area of abilities, also ignoring personality factors, physical factors, and external circumstances.

The Solution of a Three-variable Problem.—We first take the simplest case of multiple correlation, that between the dependent variable and two independent variables. In the general problem given by the data in Table 16.1, we may ask what is the correlation between freshman marks on the one hand and the two variables analogies-test scores and high-school averages on the other. The simplest general formula for this case is

$$R^2_{1.23} = \frac{r^2_{12} + r^2_{13} - 2r_{12}r_{13}r_{23}}{1 - r^2_{23}} \quad \text{(Square of coefficient of multiple correlation with three variables)} \qquad (16.1)$$

where $R_{1.23}$ = coefficient of multiple correlation between X_1 and a combination of X_2 and X_3.

Be sure to notice that this formula merely gives us R^2, the square root of which is R.

The immediate example we have set for ourselves is to find $R_{1.34}$ rather than $R_{1.23}$. To use formula (16.1), we need merely to substitute the subscripts 3 and 4 for 2 and 3. The solution is

$$R^2_{1.34} = \frac{(.583)^2 + (.546)^2 - 2(.583)(.546)(.396)}{1 - (.396)^2}$$

$$= \frac{.339889 + .298116 - .252108}{1 - .156816}$$

$$= .45766$$

$$R_{1.34} = .677$$

The Multiple-regression Equation.—We also have here a prediction problem of estimating X_1 values from both X_3 and X_4. This calls for a regression equation that involves all three variables, in other words, a multiple-regression equation. From such an equation, we can predict an X_1 value for every individual. The correlation between these predicted values (X'_1) and the obtained ones (X_1) would be .677. This is another interpretation of a multiple coefficient of correlation. For the three-variable problem, the regression equation has the general form $X'_1 = a + b_{12.3}X_2 + b_{13.2}X_3$. As in previous regression equations, the coefficient a is a constant and must be calculated from the data. Its function is to assure that the mean of the X'_1 values coincides with the mean of the X_1 values. The b coefficients serve the same purpose here as in the simple, two-variable equation. The coefficient $b_{12.3}$ is the multiplying constant or weight for the X_2 values, and $b_{13.2}$ is the weight for the X_3 values. The value of $b_{12.3}$ tells how many units X'_1 increases for every unit increase in X_2, when the effects of X_3 have been nullified or held constant. The value of $b_{13.2}$ tells how many units X'_1 increases for every unit increase in X_3, with the effects of X_2 removed from consideration. The particular b weights, as computed by the formulas given below, are the *optimal* weights. They assure the maximum correlation between predicted and obtained values. The solution, with the obtained b weights, satisfies the principle of least squares in that the sum of the squares of discrepancies between the X_1 values and the X'_1 values will be a minimum.

Solution of the b Coefficients.—We do not find the b coefficients directly from the correlations but do so indirectly through the so-called beta coefficients. Beta coefficients are called *standard partial regression coefficients*—*standard*, because they would apply if standard measures were used in all variables; *partial*, because, as in the case of the coefficient of partial correlation (see Ch. 13), the effects of other variables are held constant. The $b_{12.3}$ and $b_{13.2}$ are known as *partial regression coefficients*, because they, too, are weights that presuppose that other independent variables are held constant. They are given by the formulas

$$b_{12.3} = \left(\frac{\sigma_1}{\sigma_2}\right) \beta_{12.3} \qquad (16.2a)$$

and (Partial regression coefficients)

$$b_{13.2} = \left(\frac{\sigma_1}{\sigma_3}\right) \beta_{13.2} \qquad (16.2b)$$

The betas, in turn, are found by the formulas

$$\beta_{12.3} = \frac{r_{12} - r_{13}r_{23}}{1 - r^2_{23}} \qquad (16.3a)$$

and (Standard partial regression coefficients)

$$\beta_{13.2} = \frac{r_{13} - r_{12}r_{23}}{1 - r^2_{23}} \qquad (16.3b)$$

Similar equations apply, with change of subscripts, when the independent variables are X_3 and X_4 instead of X_2 and X_3. In our example

$$\beta_{13.4} = \frac{.583 - (.546)(.396)}{1 - (.396)^2} = \frac{.3668}{.8432} = .435$$

and

$$\beta_{14.3} = \frac{.546 - (.583)(.396)}{1 - (.396)^2} = \frac{.3151}{.8432} = .374$$

We can now solve for the b-coefficients by means of formulas (16.2ab):

$$b_{13.4} = \frac{9.1}{17.0}(.435) = .233$$

and

$$b_{14.3} = \frac{9.1}{19.4}(.374) = .175$$

For the complete regression equation, the a coefficient is still lacking. It is given by the general formula

$$a = M_1 - b_{12.3}M_2 - b_{13.2}M_3 \qquad (16.4)$$

Inserting the known values

$$a = 73.8 - (.233)(49.5) - (.175)(61.1)$$
$$= 73.8 - 11.53 - 10.69$$
$$= 51.58$$

The complete regression equation will then read

$$X'_1 = 51.58 + .233X_3 + .175X_4$$

To interpret the equation, we may say that for every unit increase in X_3, X_1 is increasing .233 unit and that for every unit increase in X_4, X_1 is increasing .175 unit. To apply the equation to a particular student whose X_3 score is 25 and whose X_4 score is 32, we predict that his X_1 score will be

$$X'_1 = 51.58 + 5.82 + 5.60 = 63.00$$

We use X'_1 to stand for his predicted average freshman mark, because he has an actual average mark that we call X_1. Some other examples of

TABLE 16.2.—SOME PREDICTIONS OF SCHOLARSHIP MARK FROM MEASURES IN TWO VARIABLES

	Student				
	A	*B*	*C*	*D*	*E*
X_3 analogies score	25	27	48	85	87
X_4 high-school average	32	61	65	90	52
$b_{13.4}X_3$	5.82	6.29	11.18	19.80	20.27
$b_{14.3}X_4$	5.60	10.68	11.38	15.75	9.10
X'_1 (predicted mark)	63.0	68.6	74.1	87.1	81.0

individual students are presented in Table 16.2 to show how various combinations of values for X_3 and X_4 point to corresponding values of X_1.

Multiple Predictions by a Graphic Method.—A graphic method of making predictions of scores in X_1 from different combinations of scores in X_3 and X_4 is shown in Fig. 16.2. The chart is drawn to apply to the prediction of average freshman grades from scores in the analogies test and high-school average. Diagonal lines are drawn in the figure, each representing the locus of identical predicted values. These lines represent X'_1 scores at intervals of 5 units. Note, for example, the line for $X'_1 = 70$. A prediction of 70 may arise from many different combina-

tions of X_3 and X_4. Choose several values, in turn, in the analogies test, for example, 10, 30, 50 and 70. Corresponding values in high-school average needed to yield predictions of 70 are 92, 65, 38, and 12, respectively. The chief use of the chart, however, is to find X'_1 for two given values in X_3 and X_4. For an X_3 of 20 and an X_4 of 50, the prediction is exactly 65. For an X_3 of 90 and an X_4 of 14, the prediction is exactly 75. When the prediction is not exactly on one of the diagonal lines, we interpo-

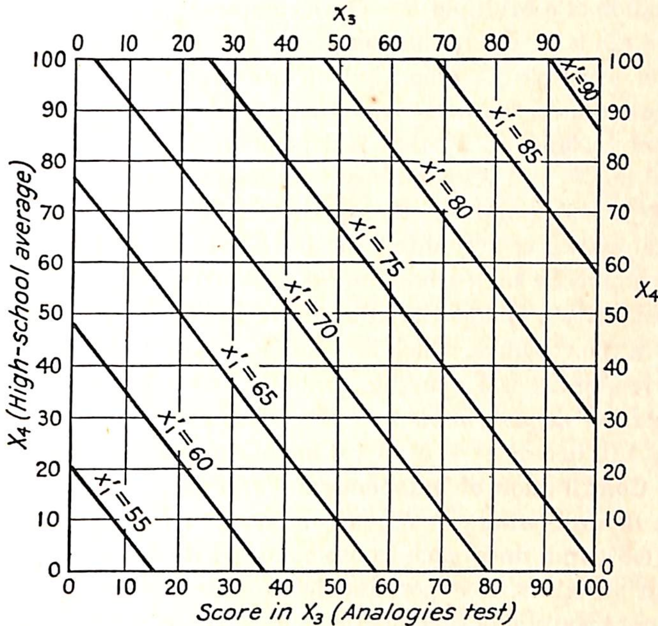


FIG. 16.2.—Diagram showing constant values in the dependent variable for different combinations of scores in two independent variables, each weighted as called for by the multiple-regression equation.

late, by inspection, between two lines. Thus, for $X_3 = 40$ and $X_4 = 70$, the most probable X_1 is 73. The proportion of the distance between two diagonal lines must be estimated by the perpendicular distance between them. The perpendicular is in a diagonal direction. The reader may get further practice in using the chart by verifying the predictions found by computation in Table 16.2.

Calculating the Multiple R from Beta Coefficients.—If the beta coefficients are known, the shortest route to the multiple R is by way of the equation

$$R^2_{1.23} = \beta_{12.3}r_{12} + \beta_{13.2}r_{13} \qquad (16.5)$$

Again, note that this gives R^2, from which the square root must be obtained. For the scholarship data and variables X_3 and X_4,

$$
\begin{aligned}
R^2_{1.34} &= (.435)(.583) + (.374)(.546) \\
&= .253605 + .204204 \\
&= .457809 \\
R_{1.34} &= .677
\end{aligned}
$$

as was found by formula (16.1) previously.

Interpretation of a Multiple R.—Once computed, a multiple R is subject to the same kinds of interpretation, as to size and importance, as were described for a simple r. One kind of interpretation is in terms of R^2, which we call the *coefficient of multiple determination*. This tells us the proportion of variance in X_1 that is dependent upon, or associated with, or predicted by X_3 and X_4 combined *with the regression weights used*. In this case, R^2 is .4578, and we can say that 45.78 per cent of the variance in freshman marks is accounted for by whatever is measured by the analogies test and by high-school marks taken together, eliminating from double consideration things that they have in common. The remaining percentage of the variance, which is 54.22 $(1 - R^2)$, is still to be accounted for. This remainder is given the symbol K^2 and is known as the *coefficient of multiple nondetermination*. This is consistent with the fact that $R^2 + K^2 = 1.0$, just as $r^2 + k^2 = 1.0$ in the simple correlation problem.

Relative Contribution of Independent Variables.—Since the coefficient of multiple determination, or R^2, is composed of the two components in formula (16.5) and since each component pertains only to one of the independent variables, it is permissible to take each component as indicating the contribution of one independent variable to the total predicted variance of X_1. This being the case, the first term, .253605, indicates the contribution to freshman scholarship by ability in the analogies test, and the second term, .204204, indicates the contribution of the high-school average. Rounded, in terms of percentages, these are 25.4 and 20.4 respectively. This enables us to obtain a more definite idea of the relative importance of each variable in the regression equation. We can say that ability in the analogies test, with what it has in common with high-school scholarship held constant, contributes about 25 per cent to freshman scholarship and that high-school marks, apart from that portion related to analogies-test ability, contribute about 20 per cent. We cannot take these as final or absolute, for there are other factors contributing to freshman scholarship level that have not been similarly eliminated from consideration. But it is of much value to be able to compare contributions of variables to outcomes in this manner.

The Standard Error of Estimate from Multiple Predictions.—The standard error of estimate is again brought in to indicate about how far the predicted values would deviate from the obtained ones. The formula is the same as previously, except that the multiple R is substituted for r. It now reads

$$\sigma_{1.23} = \sigma_1 \sqrt{1 - R^2_{1.23}} \qquad \text{(Standard error of multiple estimate)} \qquad (16.6)$$

In the illustrative problem,

$$\begin{aligned} \sigma_{1.34} &= 9.1 \sqrt{1 - .457809} \\ &= 9.1 \times .736 \\ &= 6.7 \end{aligned}$$

We can now say that two-thirds of the obtained X_1 values will lie within 6.7 points of the predicted X_1 values. The margin of error *with* knowledge of X_3 and X_4 is 73.6 per cent as great as the margin of error would be without that knowledge. These conclusions presuppose predictions made on the basis of the regression equation that was obtained, and predictions made for individuals belonging to the population and sampled at random.

The index of forecasting efficiency may also be used by way of interpretation and because of its close relation to the standard error of estimate may be mentioned at this point. The formula is the same as for a Pearson r [see formula (15.22)]. In the example of our three variables, $E = 26.4$ per cent, which means that predictions by means of the equation are 26.4 per cent better than those made merely from a knowledge of the mean of the X_1 values.

Multiple Correlation in Small Samples.—For small samples,—and for multiple-correlation problems this means anything less than an N of 100—degrees of freedom should be considered in dealing with questions of sampling. If the multiple R and the other statistics derived from it are to be used for estimating population parameters, there is even more bias than for a simple correlation problem. It was stated earlier that the multiple R represents the maximum correlation between a dependent variable and a weighted combination of independent variables. The least-square solution that is represented in computing the combined weights assures this result; but it really assures too much. It capitalizes upon any chance deviations that favor high multiple correlation. The multiple R is therefore an inflated value. It is a biased estimate of the multiple correlation in the population. If we were to apply the regression weights in a new sample and to correlate predicted X_1 values with obtained X_1 values, we would probably find that the correlation would be smaller than R. It is desirable, therefore, to find some means of estimating a parameter $\tilde{R}$

which gives a more realistic picture of the general situation. A common way of "shrinking" R to a more probable population value is by the formula

$$_cR^2 = 1 - (1 - R^2)\left(\frac{N - 1}{N - m}\right) \quad \text{(Correction in } R \text{ for bias)} \quad (16.7)$$

where N = the number of cases in the sample correlated.
m = the number of variables correlated.
$N - m$ = the number of degrees of freedom, one degree being lost for each mean, there being one mean per variable.

For the illustrative problem above, where $R = .677$, the corrected R^2 would be

$$\begin{aligned} _cR^2 &= 1 - (1 - .4579)\left(\frac{174 - 1}{174 - 3}\right) \\ &= 1 - (.5422)(1.0117) \\ &= .4515 \end{aligned}$$

from which $_cR = .672$. The correction does not make much difference here because the sample was fairly large and the number of variables small. There are problems in which the change would be very appreciable.

A similar correction is necessary for the standard error of estimate, unless $_cR$ has been used in formula (16.6), The general formula is

$$\begin{aligned} {}_c\sigma_{1.23\cdots m} &= \sigma_{1.23\cdots m}\sqrt{\frac{N - 1}{N - m}} \\ &= \sigma_1\sqrt{(1 - R^2)\frac{N - 1}{N - m}} \end{aligned} \quad \text{(General correction of a multiple standard error of estimate for bias)} \quad (16.8)$$

where the symbols N and M are as defined above. This correction also makes the greatest difference when N is small and m is large.

Sampling Errors in Multiple-correlation Problems.—Since, as was explained in the preceding section, sampling errors play a more important role in multiple-correlation problems than in simple-correlation problems, we may as well regard most samples as "small" when we consider standard errors. This means the use of degrees of freedom, and in linear relationships the number of degrees of freedom is $N - m$.

Standard Error of R.—For an R derived from any number of variables, the standard error is

$$\sigma_R = \frac{1 - R^2}{\sqrt{N - m}} \quad \text{(Standard error of a multiple } R\text{)} \quad (16.9)$$

The result is interpreted as for the standard error of any r.

When the null hypothesis is to be tested, Table D is most convenient. The R's meeting the 5 per cent and 1 per cent levels of significance are shown in columns headed by numbers of variables and rows headed by appropriate numbers of *df*. In the illustrative problem, $N = 174$, so the number of degrees of freedom is 171. The standard error of R is .041. The obtained R cannot very well be more than .11 from the population value of R (.11 being about 2.58 times σ_R). From Table D we find that with 150 degrees of freedom (the next lower and nearest to 171) and with 3 variables, an R of .198 is significant at the 5 per cent level and one of .244 at the 1 per cent level. We should have little room for doubt that a genuine multiple correlation exists in the population.

Standard Error of a Multiple-regression Coefficient.—For the beta coefficient the standard error is estimated by the formula

$$\sigma^2_{\beta_{12.34\ldots m}} = \frac{1 - R^2_{1.234\ldots m}}{(1 - R^2_{2.34\ldots m})(N - m)} \quad \text{(Standard error of a beta coefficient)} \quad (16.10)$$

The new symbol here is $R_{2.34\ldots m}$, which is a multiple R with X_2 as the dependent variable and all other variables except X_1 as independent variables. There would be one of these standard errors for each of the independent variables in turn, each being substituted for X_2. For a three-variable problem, the R in the denominator reduces to r_{23}. Note that this formula gives the *variance error*, *i.e.*, σ^2.

For the b coefficient, the standard error is estimated by

$$\sigma_{b_{12.34\ldots m}} = \frac{\sigma_{1.234\ldots m}}{\sigma_{2.34\ldots m}\sqrt{N - m}} \quad \text{(Standard error of a } b \text{ coefficient)} \quad (16.11)$$

Needed in the denominator for each independent variable in turn is the standard error of estimate of that variable from all other independent variables. Beyond a three-variable problem this becomes quite laborious, but in the latter the denominator term reduces to σ_{23}. Unlike the preceding formula, this gives the standard error *without* extracting a square root after it is solved.

The chief use of these standard errors is to test the null hypothesis, to determine whether each independent variable has anything at all to contribute to prediction when its relation to other variables is taken into account. If the obtained beta or b is not significantly different from zero, that variable might well be dropped from the regression equation, and a new equation derived. Some variables may add no more to a multiple R than would be well within the margin of error as indicated by σ_R. Rather than to go to the trouble of computing the standard errors σ_β and

σ_b, however, a decision could probably be more quickly reached, and perhaps just as dependably, by noting the proportion of variance a variable adds to R^2 and comparing this increment to R with σ_R. For example, in this simple problem already considered, we find that the analogies test taken alone could account for about 34 per cent of the variance in freshman grades; the correlation r_{13} was .583. The high-school average when taken together with the analogies test could account for about 20 per cent additional variance, bringing the correlation to .677. With a standard error for r_{13} equal to .051, it is unlikely that the correlation of .583 could have arisen by random sampling from a population in which the correlation $\tilde{r}_{13}$ is .677. With a standard error of .041 for the multiple correlation, it is unlikely that R could have arisen by random sampling from a population in which the $\tilde{R}$ is only .583. We therefore feel much confidence that X_4 has something unique to offer to predictions and something that is not merely favorable errors in the sample alone.

Some Principles of Multiple Correlation

While multiple-correlation problems may be extended to any number of variables, before we consider the solution with more than three, it is desirable to examine some of the general principles which apply for any number of variables but which can be seen more clearly when there are only three. The two main principles are (1) a multiple correlation increases as the size of correlations between dependent and independent variables increases and (2) a multiple correlation increases as the size of intercorrelations of independent variables decreases. A maximum R will be obtained when the correlations with X_1 are large and when intercorrelations of X_2, X_3, . . . , X_m are small. In building a battery of tests to predict a criterion, test makers have usually tried to maximize the validity of each test and to minimize the correlations between tests. There are limitations to the application of these objectives, however, and in practice they tend to conflict, as we shall see. There are also apparent exceptions to the rules, as examples will show. The whole story is not told by the two principles as stated.

Some Typical Combinations of r_{12}, r_{13}, and r_{23}.—Table 16.3 provides some examples of various combinations of correlations among three variables that enter into a multiple-correlation problem. The mathematically wise student will be able to predict the kind of outcome in each instance, from a general inspection of formula (16.1). Repeated here for ready reference, it reads

$$R^2_{1.23} = \frac{r^2_{12} + r^2_{13} - 2r_{12}r_{13}r_{23}}{1 - r^2_{23}}$$

If the correlation r_{23} is zero, the third term in the numerator is zero, which has a tendency to make $R_{1.23}$ larger. On the other hand, there is a distinct advantage in having r_{23} very large, because of its role in the denominator. If r_{23} approaches 1.0, the denominator approaches zero. Even

TABLE 16.3.—EXAMPLES OF MULTIPLE CORRELATIONS IN A THREE-VARIABLE PROBLEM WHEN INTERCORRELATIONS VARY

Example	r_{12}	r_{13}	r_{23}	$R^2_{1.23}$	$R_{1.23}$
1	.4	.4	.0	.3200	.57
2	.4	.4	.4	.2286	.48
3	.4	.4	.9	.1684	.41
4	.4	.2	.0	.2000	.45
5	.4	.2	.4	.1619	.40
6	.4	.2	.9	.2947	.54
7	.4	.0	.0	.1600	.40
8	.4	.0	.4	.1905	.44
9	.4	.0	.9	.8421	.92
10	.4	.2	−.4	.3143	.56
11	.4	−.4	−.4	.2286	.48

though the numerator may become small, under these conditions R could be quite large. A large R is thus favored by having r_{23} either very small or very large. This principle should be added to the two mentioned above. But it should be said also that a large r_{23} is more effective when the independent variables are unequally correlated with the dependent variable, and particularly when one of the correlations is very small.

Note the first example in Table 16.3, in which $r_{23} = .0$. For this event, formula (16.1) reduces to

$$R^2_{1.23} = r^2_{12} + r^2_{13} \quad \text{(Multiple } R \text{ when intercorrelations of two independent variables are zero)} \qquad (16.12)$$

In other words, when independent variables are not correlated, the proportion of variance predicted by their combination is equal to the sum of the proportions of variance predicted by each separately. This holds for any number of independent variables whose intercorrelations are zero. A psychological interpretation of this is that when intercorrelations among predictive measures are zero, the total contribution of each to the prediction of a complex criterion containing all the things predicted is unique.

Note next the second and third examples and compare them with the first. In all three, the r_{12} and r_{13} correlations remain constant at .4, while

r_{23} increases first to .4 then to .9. As this happens, R goes from .57 to .48 to .41. In the last instance r_{23} is so high that there is practically no gain from combining the two variables X_2 and X_3. We shall see a modified result in the next three examples.

In examples 4 to 6, r_{12} remains constant at .4 and r_{13} constant at .2, while r_{23} varies from .0 to .9. In the first of these three we find formula (16.12) verified. The two variances sum to .2000 and R is .45. As r_{23} increases to .4, R shrinks back to approximately .40. Thus we can conclude that if one test has a validity of .4, it may pay to add to it another with a validity of only .2 provided the two tests intercorrelate zero. But if there is any appreciable correlation between them, or only a moderate correlation, it would not pay. What happens if we increase r_{23} still more? When it is as high as .9, R jumps to .54. This supports the third principle stated above: that r_{23} should be either very low or very high. One may ask why this principle did not appear to work in the first three examples. The answer is that it was obscured by the relation of r_{12} and r_{13}. In those examples r_{12} equaled r_{13}, and in the next three examples these correlations were unequal. A better explanation is that one of them is very small. One may well ask what psychological meaning is involved in the increase in R when r_{23} is very large. This is best explained in connection with the next three examples.

In examples 7 to 9, r_{12} and r_{13} are still more uneven in size. They also have special interest because $r_{13} = .0$ in all three, while r_{23} varies from .0 to .4 to .9, as in the previous groups of three examples. It would seem, at first thought, that any test that correlates zero with a criterion would have no value in predicting that criterion. It is true that alone it has no value whatever for doing so. But it is not true if that test is combined with other tests with which it correlates. In example 7, the common-sense expectation is vindicated. The addition of an invalid test would offer no improvement. It would simply receive a regression weight of zero which means it would not be included in the regression equation. But note that when r_{23} is increased to .4, R becomes .44 and when r_{23} is .9, R becomes .92. Clearly a test with zero validity may add materially to prediction if it correlates substantially with another test that *is valid.*

Suppression Variables.—The psychological significance of this is best explained by factor theory (see Ch. 18). Roughly, the answer is that variable X_2, in spite of its positive correlation with X_1, has some variance in it that correlates either zero, or perhaps even negatively, with the criterion. This same variance prevents X_2 from correlating as highly as it might with X_1. Variable X_2 correlates with X_3 because they have in common that variance not shared by X_1. In this kind of situation we

find that X_3 acquires a *negative* regression weight, although it may correlate only zero, and not negatively, with the criterion. We call such a variable a *suppression variable.* Its function in a regression equation is to suppress whatever variance in other independent variables may not be represented in the criterion but which may be in some variable that does otherwise correlate with the criterion.

An example of this came to the author's attention in testing for pilot selection. It was a consistent finding that a vocabulary test, which is as pure a measure of the verbal-comprehension factor as we have, correlated zero or even slightly negative with the criterion of success in pilot training. The same kind of test correlated substantially with a reading-comprehension test which also correlated positively with the pilot criterion. The reading test correlated positively with the criterion because it measured, besides verbal comprehension, such factors as mechanical experience and visualization which were also component variances in the pilot criterion. The combination of a vocabulary test with the reading test, with a negative weight for the vocabulary test, would have improved predictions over those possible with the reading test alone.

The examples mentioned thus far have had only positive correlations involved. In most practice where human variables are measured we have only zero or positive correlations, if all measurement scales are aligned so that "good" qualities are given high numerical values. Where genuinely negative relationships do occur they are likely to be very small. Examples 10 and 11 in Table 16.3 are given more for their academic than for their practical interest. Example 10 should be compared with examples 4, 5, and 6. They differ only in the value of r_{23}. When r_{23} becomes negative, we see that the increase that occurred when r_{23} approaches zero appears to continue as r_{23} becomes increasingly negative. When r_{23} is $-.4$, R is even greater than when r_{23} is .9. It is doubtful whether situations like example 10 occur in nature, though they are theoretically possible. The trend could not go too far, however, for with r_{23} large enough in the negative direction we would come to a multiple R greater than 1.0, which would mean an impossible situation, even mathematically.

Example 11 has two negative correlations, r_{13} and r_{23}. These simply mean that variable X_3 probably has a reversed scale, for X_3 is related to both X_1 and X_2 in the same direction. Note that the multiple R is the same as if both r_{13} and r_{23} were positive and of the same size numerically (example 2).

Multiple-*R* Principles in Larger Batteries.—The principles illustrated above for the three-variable problems also apply in larger combinations of variables. The first two principles can be well illustrated by taking other

hypothetical examples like those in Table 16.4. There we have a demonstration of how multiple R's behave as the number of independent variables increases from 2 to 20 and as intercorrelations increase from .0 to .6.

TABLE 16.4.—MULTIPLE CORRELATIONS FROM DIFFERENT NUMBERS OF INDEPENDENT VARIABLES EACH CORRELATING .30 WITH THE DEPENDENT VARIABLE BUT WITH INTERCORRELATIONS VARYING*

Number of independent variables	Intercorrelations			
	.00	.10	.30	.60
1	(.30)	(.30)	(.30)	(.30)
2	.42	.40	.37	.34
4	.60	.53	.44	.36
9	.90	.67	.48	.37
20	—	.79	.52	.38

* Adapted from Thorndike, R. L. Research problems and techniques, in the AAF Aviation psychology research program reports, No. 3. Washington, D.C.: Government Printing Office, 1947.

Following Thorndike's choices, we will assume that each variable correlates with a criterion to the extent of .3. This is a rather low validity coefficient, and about the lower limit of usefulness for a single test or other predictive device. We will see, however, how valuable such instruments may be when combined in a battery, provided their intercorrelations are not too high.

In the second row of Table 16.4, when two such tests are combined, we see how the multiple R decreases from .42 when r_{23} is zero to .34 when r_{23} is .60. In each row the same expected phenomenon occurs: a decrease in R as intercorrelations increase. Inspection of the columns shows how R increases as we add more tests of the same kind to the battery and how the gain in R continues up to a battery of 20, except for the case of zero intercorrelations, for which the limit of $R = 1.0$ was passed when the number of tests exceeded 11. In this situation (intercorrelations zero) the principle of formula (16.12) still applies. The proportion of predicted variance contributed by each test would be .09, and 11 tests would yield a multiple R of .99. In other columns the increases of R are less drastic, but except in the last column, and perhaps in the one preceding, it would apparently pay to continue adding new tests until the 20 were included. Matters of administrative effort would have to be balanced against gains in R.

Table 16.4 tells an even more important story. The value of having zero intercorrelation among tests in a battery is obvious. If one tries to achieve zero intercorrelations among tests, each test measuring a unique

factor, however, he will often find that each test tends to correlate low with the criterion. This is because a practical criterion, of training achievement or of job performance, is usually a complex variable; it has a number of component variances, each component being a common factor (see Ch. 18). If one tries to increase the correlation of a test with a criterion, the result is almost invariably to increase the factorial complexity of the test; to bring in more different factor variances. This automatically raises the correlation of this test with other tests, because they have more factors in common. This is the reason that in practice the two principles mentioned first lead to conflicting objectives. Where there has to be a choice, it seems wisest to give less attention to the first principle (of maximizing correlation of each test with the criterion) and greater attention to the second (of minimizing intercorrelations). If there are 20 independent factors represented in a practical criterion, and if each is of equal importance, each would contribute .05 of the total variance. Each test, measuring only one of the factors, would need to correlate only $\sqrt{.05}$, which is .224, with the criterion. In this case, raising the correlation between any one test and the criterion would be of little use. There would be no objection to a higher correlation. Appropriate weighting would bring the test's contribution to prediction down to required proportions. Thus, it can be concluded that low correlations of tests with practical criteria can be tolerated provided we can combine enough tests in a battery and provided their intercorrelations are near zero.[1]

Multiple Correlation with More than Three Variables

With more than three variables, the best solution of a regression equation and of a multiple R is by means of the Doolittle method. This procedure will be outlined step by step for a five-variable problem. We shall use all the variables represented in Table 16.1, asking what regression weights would best predict X_1 from the other four combined and what the correlation of those predictions with obtained X_1 values would be.

Solution of Normal Equations.—The mathematically inclined reader will appreciate better what is transpiring in applying the Doolittle method if he knows that he is actually solving simultaneous equations. The unknowns are the beta coefficients, and there are as many equations as unknowns. For a five-variable problem, in which there are four unknown betas, the equations are

[1] For a more detailed discussion of these problems, see Guilford, J. P. New standards for test evaluation. *Educ. & Psychol. Meas.*, 1946. **6**, 427–438.

$$\begin{aligned} \beta_{12} + r_{23}\beta_{13} + r_{24}\beta_{14} + r_{25}\beta_{15} &= r_{12} \\ r_{23}\beta_{12} + \beta_{13} + r_{34}\beta_{14} + r_{35}\beta_{15} &= r_{13} \\ r_{24}\beta_{12} + r_{34}\beta_{13} + \beta_{14} + r_{45}\beta_{15} &= r_{14} \\ r_{25}\beta_{12} + r_{35}\beta_{13} + r_{45}\beta_{14} + \beta_{15} &= r_{15} \end{aligned} \quad \text{(Normal equations for the solution of beta weights)} \quad (16.13)$$

The beta coefficients are symbolized in abbreviated form here to conserve space. β_{12}, in full, would be $\beta_{12.345}$ and β_{13} would be $\beta_{13.245}$, and so on. The equations are systematic, the r coefficients being arranged as in the original table of intercorrelations (see Table 16.1). The betas in the diagonal positions might be expected to have coefficients r_{22}, r_{33}, r_{44}, and r_{55} attached to them, but instead the coefficients attached to these betas are all +1.0, as the least-square solution requires.

The Doolittle-solution Operations.—First we prepare a work sheet like that in Table 16.5. There is a column for every variable and the number-

TABLE 16.5.—SOLUTION OF A MULTIPLE-CORRELATION PROBLEM BY THE DOOLITTLE METHOD

Column number		2	3	4	5	1	Check
Variable		X_2	X_3	X_4	X_5	X_1	Sum
Row	Instruction						
A	r_{2k}	1.0000	.5620	.4010	.1970	.4560	2.6250
B	$A \div (-A2)$	−1.0000	− **.5620**	− **.4010**	− **.1970**	− **.4560**	−2.6250
C	r_{3k}	—	1.0000	.3960	.2150	.5830	2.7560
D	$A \times B3$	—	− .3158	− .2254	− .1107	− .2613	−1.4752
E	$C + D$	—	.6842	.1706	.1043	.3217	1.2808
F	$E \div (-E3)$	—	−1.0000	− **.2493**	− **.1524**	− **.4702**	−1.8720
G	r_{4k}	—		1.0000	.3450	.5460	2.6880
H	$A \times B4$	—		− .1608	− .0790	− .1865	−1.0526
I	$E \times F4$	—		− .0425	− .0260	− .0802	− .3193
J	$G + H + I$	—		.7967	.2400	.2793	1.3161
K	$J \div (-J4)$	—		−1.0000	− **.3012**	− **.3506**	−1.6519
L	r_{5k}	—			1.0000	.3650	2.1220
M	$A \times B5$	—			− .0388	− .0916	− .5171
N	$E \times F5$	—			− .0159	− .0490	− .1952
O	$J \times K5$	—			− .0723	− .0841	− .3964
P	$L + M + N + O$	—			.8730	.1403	1.0133
Q	$P \div (-P5)$	—			−1.0000	− **.1607**	−1.1607

ing corresponds. A last column is introduced for the purpose of checking the calculations, as will be explained. The rows are designated by letters, and in the first column, a shorthand instruction is noted. These will be explained.

Step 1. Record in row A the correlations with X_2. These are obtained here from Table 16.1. In column (2), a coefficient of 1.0000 is inserted, because it is demanded by the Doolittle method. We are going to carry four decimal places throughout the solution (one more than those given in the r's); so we record all numbers to four places.

Step 2. Sum the values recorded in row A, and give the sum in the last or "check" column. This will be used later.

Step 3. Divide the numbers in row A each by -1.0000. In the table, the instruction reads "$A \div (-A2)$," which means that each number in row A is to be divided by the number that appears at $A2$ [row A, column (2)] with sign changed. This includes the last column as well.

Step 4. Record in row C all the remaining correlations with X_3. We say "remaining," because one is already recorded, namely, r_{23}. The value of 1.0000 is recorded at $C3$.

Step 5. Sum all the correlations with X_3, including the .5620 in row A. Record the sum in the "check" column.

Step 6. The numbers in row D are found by the instruction "$A \times B3$," which means to multiply all the numbers in row A [beginning in column (3)] by the number that appears in row B and column (3). This number is $-.5620$ in Table 16.5.

Step 7. Row E calls for the addition of all numbers in rows C and D.

Step 8. Row F calls for the division of all numbers in row E by the number appearing in row E and column (3), with sign changed. This number, with sign changed is $-.6842$.

Step 9. We are ready for the first checking of calculations. Sum the values in row F, *not* including the last column. This should equal approximately -1.8720 in this particular problem, which was found by the steps already described. If there is a serious discrepancy here (other than in the fourth decimal place), check row E by adding values up to the check column. If this does not check, there is an error further back, and some recalculating is in order. All checks should be satisfied before proceeding.

Step 10. In row G, record remaining correlations with X_4, with 1.0000 at $G4$.

Step 11. Sum *all* the correlations with X_4, and record in the last column in row G.

Step 12. Values in row H are the products of values in row A times the number at $B4$. This number is $-.4010$.

Step 13. Values in row I are the products of numbers in row E times the number at $F4$, which is $-.2493$.

Step 14. Sum the numbers in rows G, H, and I for each column.

Step 15. Divide row J through by the number at $J4$, with sign changed; in other words, by $-.7967$.

Step 16. Check by summing row K up to the last column. Does the sum agree with the number already found in that column?

Step 17 and after. By now the abbreviated instructions for each row should be clear by analogy to those already given. The final check is made in row (Q).

The illustrative solution is set up for a five-variable problem, but a larger number of variables would be treated in a similar manner simply by extending the table to more rows and columns. A smaller number of variables would mean fewer rows and columns. It will be noticed that the table is set up in terms of *blocks* of work, each one beginning with the entrance of correlations for a new variable and ending by dividing by a number that will assure a -1.0000 as the first number in the last row of that block. The work will be found to be very systematic throughout. Any variable may be treated as the dependent variable, but it must then occupy the next to the last column in the table.

Solution of the Beta Coefficients. The work represented in Table 16.5 is only a part of the Doolittle solution. The end result gives the beta coefficients, which we find by means of a "back solution," so called because we work in a backward direction, as compared with the work in Table 16.5. This work can be tabulated, but it is probably clearest to the beginner in the form of equations. The first beta found is β_{15}, which can be located without further ado in Table 16.5. It is the number at the intersection of row Q and column (1), but with sign changed (in other words, it is described as $-Q1$). β_{15} is therefore $+.1607$. The other betas require more work; so we will follow the procedure step by step, including again the first step already taken, for the sake of completeness.

Step 1. $\beta_{15} = -Q1 = +.1607$

Step 2. $\beta_{14} = -K1 + \beta_{15}(K5) = .3506 + (.1607)(-.3012) = +.3022$

Step 3. $\beta_{13} = -F1 + \beta_{15}(F5) + \beta_{14}(F4)$

$= .4702 + (.1607)(-.1524) + (.3022)(-.2493) = +.3703$

Step 4. $\beta_{12} = -B1 + \beta_{15}(B5) + \beta_{14}(B4) + \beta_{13}(B3)$
$= .4650 + (.1607)(-.1970) + (.3022)(-.4010) + (.3703)(-.5620)$
$= +.1039$

Before going further, it is well to check the calculations of the beta coefficients. This can be done by using the last equation in (16.13):

$$\beta_{12}r_{25} + \beta_{13}r_{35} + \beta_{14}r_{45} + \beta_{15} = r_{15}$$

Substituting known values,

$$(.1039)(.197) + (.3703)(.215) + (.3022)(.345) + .1607 = .3651$$

Since $r_{15} = .365$, the check is satisfied, and we may assume that there has been no error in computing the betas. This checking procedure can be summarized as in Table 16.6, which provides a convenient work plan.

TABLE 16.6.—A CHECK UPON THE COMPUTATION OF THE BETA COEFFICIENTS

	β_{1k}	r_{k5}	$\beta_{1k}r_{k5}$
X_2	.1039	.197	.0205
X_3	.3703	.215	.0796
X_4	.3022	.345	.1043
X_5	.1607	1.000	.1607
			Σ .3651 = r_{15}

The Solution of Regression Weights and the Multiple *R*. Each b coefficient needed in the multiple-regression equation is found from its corresponding beta. Equations like those in formulas (16.2*a*) and (16.2*b*) apply. The b weight for X_2 should now read in full $b_{12.345}$ to indicate that we are interested in the relation of X_1 to X_2, other variables, X_3, X_4, and X_5, being held constant. For the sake of brevity (as, indeed, we have already done for the betas), we shall denote the b's only by the first two subscript numbers b_{12}, b_{13}, etc. In the solution of a multiple R, equation (16.5) needs to be extended to include as many terms as there are variables. R^2 is the sum of the products of beta times its corresponding r, *i.e.*,

$$R^2 = \beta_{12}r_{12} + \beta_{13}r_{13} + \beta_{14}r_{14} + \beta_{15}r_{15} + \cdots \quad (16.14)$$

(General solution of R from beta coefficients)

The a coefficient in the equation is also found by formula (16.4), extended with as many terms as necessary. It is the mean of the X_1 values minus the products of other means times their corresponding b weights, as

$$a = M_1 - b_{12}M_2 - b_{13}M_3 - b_{14}M_4 - \cdots \quad (16.15)$$

(Constant a in a multiple regression equation)

All these operations are conveniently carried out in a work sheet like Table 16.7, where R and the regression weights are systematically cal-

TABLE 16.7.—SOLUTION OF THE REGRESSION COEFFICIENTS FOR THE MULTIPLE-REGRESSION EQUATION

(1)	(2) β_{1k}	(3) r_{1k}	(4) $\beta_{1k}r_{1k}$	(5) σ_1/σ_k	(6) b_{1k}	(7) M_k	(8) $(-M_k)b_{1k}$
X_2	.1039	.465	.048314	1.750	.182	19.7	− 3.585
X_3	.3703	.583	.214885	.535	.198	49.5	− 9.801
X_4	.3022	.546	.165001	.469	.142	61.1	− 8.676
X_5	.1607	.365	.058655	2.459	.395	29.7	−11.732
			Σ .487855 = R^2				Σ −33.794
			.698 = R				M_1 73.800
							a = 40.006

culated. The second column contains the four betas. The third contains the original or raw correlations of the four variables with X_1. The subscript k stands for variables 2 to 5 in turn. The fourth column contains the cross products of betas times corresponding r's. Their sum is R^2, which here is .487855; and by taking the square root we find R is .698. This R, with full subscript, would read $R_{1.2345}$.

So much for the multiple R, which we see is not increased very much by including two more variables (X_2 and X_5) over that obtained when we used only X_3 and X_4. Then R equaled .677. The coefficient of determination is now .4879, or we have accounted for 48.8 per cent of the variance of freshman scholarship, as compared with 45.8 per cent without using X_2 and X_5. The standard error of estimate (now designated as $\sigma_{1.2345}$ in full) equals 6.5, where before it was 6.7, a trifling change. The index of forecasting efficiency is now 28.4 per cent, where before it was 26.4 per cent. It is therefore questionable whether the trouble of measuring and using in the regression equation the two additional variables is worth while. This is a good example of the way in which each additional variable yields diminishing returns in the way of improved predictions.

For the solution of the b coefficients, we introduce in Table 16.7 first the column headed σ_1/σ_k. This is the ratio by which each beta is to be multiplied. The b coefficients follow in column (6). They tell how many units X_1 is increasing for each unit of increase in the other variables. From these taken alone, it would seem that X_5 (interests) has the greatest bearing upon freshman marks and that X_2 (high-school average) has the least. But such is not the situation. The best comparison of each variable's contribution to the variance in X_1 is to be seen in column (4), where each beta

is multiplied by the corresponding raw r. Here it is seen that X_3 contributes about 21 per cent, X_4 nearly 17 per cent, whereas X_5 contributes only about 6 per cent, and X_2 about 5 per cent. These statements are relative to this correlational situation, with the influences of overlapping among the four taken into account. But as to choices among the four variables that we have here, they come in the same rank order as the βr products.

For the solution of the a coefficient, the last two columns are included. This coefficient turns out to be exactly 40.0. The entire regression equation now reads

$$X'_1 = 40.0 + .182X_2 + .198X_3 + .142X_4 + .395X_5$$

With this equation, we could predict an X'_1 for every student, knowing his four scores in the other variables. As was said before, the addition of the terms involving X_2 and X_5 yield scarcely enough additional accuracy of prediction to justify their inclusion. One could try combinations of three predictive indices, variables X_2, X_3, and X_4, or X_3, X_4, and X_5, to see what happens. From the results in Table 16.7, it would seem that the last-mentioned combination of three is the more promising. One could determine by another Doolittle solution whether it increased R sufficiently above .677 to justify the inclusion of X_5 with X_3 and X_4.

Short Solutions for Regression Weights

Solution of a multiple-regression problem, even with the convenient Doolittle procedure, becomes energy and time consuming when the number of variables is large. The author has known of test batteries involving as many as 20 possible scores that could be combined each with its appropriate weight. When there are more than six variables the situation calls for possible short cuts or approximation methods. Two methods will be mentioned to meet this need, one of which will be illustrated.

The Wherry-Doolittle Method.—In recent years a modified Doolittle solution has been introduced by Wherry.[1] The method was designed to meet the requirement of assembling a battery of tests to select personnel for some particular assignment. It takes particular cognizance of the fact that when a large number of tests are validated singly for the prediction of a certain criterion, only four or five when combined often seem sufficient. As a matter of fact, adding tests beyond the point at which all the factors that the tests measure in common with the criterion are covered often merely contributes error variance to the composite. Even

[1] Described in full in Stead, W. H., Shartle, C. L., et al. Occupational counseling techniques. New York: American Book, 1940, 245–255.

before the point has been reached where there is no *apparent* improvement in prediction, errors have entered into the picture to help determine the regression weights. This point was mentioned earlier in connection with the discussion of shrinkage formulas [see formulas (16.7) and (16.8)].

The principles of the Wherry-Doolittle method are, briefly, as follows. One starts with the single test that seems to offer most in prediction of the criterion. The method then aids in selection of the second test that will have most to add to prediction when combined with the first. A third can be selected which will add most by way of prediction when combined with the first two, and so on. At each step a shrinkage formula is applied in order to determine whether the shrunken R is appreciably larger than the previous R. At the point where no further gain according to these standards is apparent, no more tests are added.

The method does undoubtedly offer an efficient way of assembling a battery of tests to meet a particular purpose. It results in a list of predictive instruments that, out of a larger number tried experimentally, is minimum for doing the job. The author is inclined toward a quite different philosophy of development of test batteries, however, which would render the Wherry-Doolittle procedure unnecessary when there is sufficient information about the criterion and the tests.[1] For this reason the space that it would take to explain and demonstrate the Wherry-Doolittle method is not used here. The reason why only four or five tests have seemed to be the limit in a useful battery is because only a limited number of the human abilities and other traits that are involved in a practical criterion have been represented in the tests. Although a dozen different tests may have been tried out, the same limited number of fundamental factors have been measured by them and the measurement is duplicated several times over. If a careful study of the criterion is made revealing *all* the factors that are worth trying to predict and if there is sufficient variety in the tests to take care of all the factors, it will be found that more than four or five tests will probably be needed. If one knows that there are 10 traits in the criterion that are worth covering with tests, and if it takes 10 tests to do it, then one could put the 10 tests in a battery and expect that every one would have something to contribute toward prediction. A successive selection of tests by a method such as the Wherry-Doolittle would then be unnecessary.

An Iterative Solution of Regression Weights.—The iterative procedure for computing beta weights to be described and illustrated is economical, particularly for a problem with many variables, and will probably lead to

[1] For a discussion of this at some length, see Guilford, J. P. Factor analysis in a test development program. *Psychol. Rev.*, 1948, **55**, 79–94.

TABLE 16.8.—AN ITERATIVE SOLUTION OF THE BETA COEFFICIENTS

	Correlations				Discrepancies ($r'_{1k} - r_{1k}$)													
	r_{2k}	r_{3k}	r_{4k}	r_{5k}	r_{1k}	r'_{1k}	d_1	d_2	d_3	d_4	d_5	d_6	d_7	d_8	d_9	d_{10}	d_{11}	d_{12}
X_2	1.000	.562	.401	.197	.465	.5283	+.0633	+.0033	+.0258	+.0179	−.0021	+.0091	−.0009	+.0047	−.0003	+.0005	+.0007	−.0003
X_3	.562	1.000	.396	.215	.583	.5742	−.0088	−.0425	−.0025	−.0111	−.0223	−.0023	−.0079	+.0022	−.0006	+.0005	+.0007	−.0002
X_4	.401	.396	1.000	.345	.546	.4680	+.0220	−.0021	+.0137	−.0001	−.0081	−.0002	−.0042	−.0002	−.0022	−.0002	+.0001	−.0003
X_5	.197	.215	.345	1.000	.365	.4074	+.0424	+.0306	+.0392	−.0008	−.0047	−.0004	−.0024	−.0003	−.0013	−.0006	+.0004	+.0004
Σr_{ak}	2.160	2.173	2.142	1.757														

Trial Betas

Trial	β'_{12}	β'_{13}	β'_{14}	β'_{15}	Trial	β'_{12}	β'_{13}	β'_{14}	β'_{15}
1	.2	.3	.3	.2	7	.11	.36	.3	.16
2	.14	.3	.3	.2	8	.11	.37	.3	.16
3	.14	.34	.3	.2	9	.105	.37	.3	.16
4	.14	.34	.3	.16	10	.105	.37	.302	.16
5	.12	.34	.3	.16	11	.105	.37	.302	.161
6	.12	.36	.3	.16	12	.104	.37	.302	.161
						.104	.370	.302	.161
						β_{12}	β_{13}	β_{14}	β_{15}

satisfactory results in most cases.[1] The operations will be described step by step and illustrated in Table 16.8 with the use of the same data to which the Doolittle method was applied earlier.

The general principle of the method is (1) to guess what the betas are going to be, (2) to substitute them in the normal equations (see equations 16.13), (3) see how much discrepancy there is between the known validity coefficients and those that follow from the guessed betas, and (4) make corrections in the guessed betas. These steps are repeated until the discrepancies practically vanish. The correlations that enter into the normal equations are listed first in the worktable, upper left-hand corner. From here on the steps will be listed.

Step 1. Compute the sum of each column of correlations, Σr_{ak}, where a stands for each of the independent variables representing columns and k stands for each of the variables in rows in turn. Σr_{ak} in the first column of correlations is Σr_{2k}, and so on.

Step 2. Make a guess for the size of each beta (these will be β'_{12}, β'_{13}, and so on) by dividing the validity coefficient for each test by the sum of its column of r's. These may be made to two decimal places to start with, but one place will do about as well. For example, β'_{12} is estimated by the ratio .465/2.160 which equals .215, but this has been rounded to .2. β'_{13} is estimated by the ratio .583/2.173, which is .268, rounded to .3. With more variables, a multiple of each such ratio would be a better estimate.

Step 3. Solve each equation, substituting the guessed betas for the unknown betas. The first equation would read

$$(1.000)(.2) + (.562)(.3) + (.401)(.3) + (.197)(.2) = .5283$$

This gives a value symbolized by r'_{1k} (for the first equation it is r'_{12}) and recorded in the column just after the validity coefficients, r_{1k}. Four decimal places will be carried from here on in order to obtain three significant digits in the betas.

Step 4. Find the discrepancy between each validity estimated from the use of the guessed betas and the obtained validity. Call these values d_1. For the first test, $d_1 = r'_{12} - r_{12} = +.0633$. This means that with the betas which were assumed, the validity of variable X_2 would have to be .0633 higher than the validity of

[1] The procedure is the author's version of R. L. Thorndike's adaptation of one originally developed by Kelley and Salisbury. See Thorndike, R. L. Research problems and techniques. AAF aviation psychology research program, Report No. 3. Washington, D.C.: Government Printing Office, 1947; also Kelley, T. L., and Salisbury, F. S. An iteration method for determining multiple correlation constants. *J. Amer. statis. Ass.*, 1926, **21**, 282*ff*.

.465 which had been obtained. The d_1 of $-.0088$ for variable X_3 indicates that the guessed betas underestimate the validity of that test.

Step 5. Make the first change in the guessed betas. Although we can see that the betas for X_2, X_4, and X_5 have been perhaps overestimated and that for X_3 underestimated, it is most convenient, and perhaps just as expedient, to make only one change at a time. Note where the largest discrepancy is. It is the $+.0633$ for variable X_2. If we make a change only in β'_{12}, it will affect only the first term in each equation and will involve only the first column of correlations. To lower d_1 to zero for the first test in the list, we would need to multiply 1.000 by some amount that will cancel it. A change of $-.0633$ would do this, but it is best to limit adjustments to the second decimal place at this stage. We will therefore reduce β'_{12} by $-.06$, making it .14.

Step 6. Modify the discrepancies in line with the change in β'_{12} just mentioned. Every d_1 will be altered by adding to it the product of the *change* times the corresponding value r_{2k}. The first d_2 will be $+.0633 + (-.06)(1.000) = +.0033$. The second d_2 will be $-.0088 + (-.06)(.562) = -.0425$, and so on.

The general pattern of the procedure is now complete. We keep on making successive adjustments as called for, computing the altered discrepancies, with an attempt to reduce them almost to zero. Since we are expecting three-place accuracy in the betas, we will find that it pays to continue until the discrepancies are not over .0005. After we have achieved good adjustment up to the second decimal place in the betas, we then proceed to make adjustments in the third decimal place. A comparison of the betas found in Table 16.8 with those found by the Doolittle solution (see Table 16.7) will show very good agreement to the third decimal place.

From the beta coefficients found in this manner one may proceed to compute the multiple correlation, the b weights, and other derived statistics.

Great care should be taken for accuracy of computation. Errors may creep in at any stage and it still might be possible to reach what looks like a satisfactory solution, that is, with zero discrepancies, with wrong betas. It would certainly be well to check the accuracy of the obtained betas as was done following the Doolittle solution. There may be some problems, with peculiar combinations of correlations, in which the iteration would not achieve zero discrepancies even after a long series of trials. The author has not encountered such a situation as yet. The routine described above

may be modified as the user of it gains experience. There are opportunities for making wiser choices of betas and changes in betas that might cut the number of steps.

Thorndike makes some suggestions concerning the original source of guessed betas.[1] If we have prior knowledge of how a given test has performed in a similar battery for making a similar prediction, it would be well to start with that knowledge. If the battery is a very large one (10 or more) it would be desirable to start with about half of the guessed betas equal to zero. Kelley and Salisbury had suggested that each beta be guessed as about half the corresponding validity coefficient, but Thorndike suggests between one-fourth and one-half is better. If a test correlates relatively low with others, the chances are that its beta will go higher than original estimates, and, conversely, if it correlates relatively high with other tests, its beta will prove to be lower.

Combinations of Measures

The regression equation is a means of combining different measures of the same object in order to derive a composite measure or score. The scores are summed, each weighted by its regression coefficient. There are other ways of combining scores to form a composite. For example, one might simply sum the raw scores for each person without applying differential weights. This is the common practice in deriving total scores of tests composed of subtests of different kinds, though in some cases there is some effort at weighting, *e.g.*, multiplying one score by 2, another by 3, and so on. Actually, every test that is composed of items may be regarded as a *battery* of as many tests as there are items. The total score is usually an unweighted summation of the item scores, though in many interest and temperament tests there may be differential weighting. Rarely does a test maker resort to the determination of regression weights for test items, but the same principle that applies to test batteries could be adapted to single tests composed of parts. More often than not, even in the case of test batteries, there are so many parts, or they are used to predict in such a variety of situations, that there is not sufficient incentive to work out the regression weights.

Because there must be substitute weighting procedures in combining tests, it is important to know some of the better substitute procedures for the multiple-regression equation and to be able to evaluate the effectiveness of a composite derived by any method. The multiple R applies only when the optimal regression weights are used; other weights will yield a composite that is likely to correlate less with the criterion. There are

[1] Thorndike, *op. cit.*

other problems connected with composite scores that call for attention, including what mean and what standard deviation will result when measures are combined each with a certain weight. These problems will be dealt with in following paragraphs.

Means of Weighted Composites.—When several measures of the same object are summed, each with its own weight, the mean of the same kind of composite for a sample of objects is given by the equation[1]

$$M_{ws} = \Sigma w_i M_i \qquad \text{(Mean of a sum of weighted measures)} \qquad (16.16)$$

where w_i = weight applied to each variable X_i, when i varies from 1 to n in a list of n variables.

M_i = mean for the same sample of objects in variable X_i.

If we apply this to the b weights computed for the regression equation in the prediction of freshmen average grades (see p. 446), the solution would be

$$M_{ws} = (.182)(19.7) + (.198)(49.5) + (.142)(61.1) + (.395)(29.7)$$
$$= 33.8$$

Thus, the mean of the composite of four variables, including X_2 (arithmetic test), X_3 (analogies test), X_4 (high-school average), and X_5 (interest score), weighted with the coefficients .182, .198, .142, and .395, respectively, would be 33.8. This value is 40.0 units short of the mean for the criterion (freshman grades). By adding the difference (40.0) which is the a coefficient of the complete regression equation, we obtain a composite mean that coincides with that of the criterion. This discussion, in other words, explains the need for the a coefficient in the complete regression equation. If we were not interested in achieving that mean, we could drop the constant 40.0 and be left with a mean of 33.8.

Standard Deviations of Weighted Composites.—We can likewise estimate the standard deviation of a composite measure when each component has a multiplier or weight. The computation of this statistic may be clearer, however if we consider the standard deviation of a simple unweighted sum first.

The Standard Deviation of Sums When Weights Equal One.—When scores from different tests are summed without applying differential weights, we may regard the weight for each test to be +1. When *two* scores are summed to make the composite, the variance of the composite scores is given by the equation[1]

[1] For proof, see Appendix A.

$$\sigma^2_s = \sigma^2_1 + \sigma^2_2 + 2r_{12}\sigma_1\sigma_2 \quad \text{(Variance of a sum of two unweighted measures)} \qquad (16.17)$$

where σ^2_1 and σ^2_2 = variances of the components.
r_{12} = coefficient of correlation between the two components.

The expression $r_{12}\sigma_1\sigma_2$ is known as the *covariance* of the two components. Its relation to correlation can be better shown by relating it to the Pearson formula, in which

$$r_{12} = \frac{\Sigma x_1 x_2}{N\sigma_1\sigma_2}$$

If we multiply both sides of this equation by $\sigma_1\sigma_2$ we have

$$r_{12}\sigma_1\sigma_2 = \frac{\Sigma x_1 x_2}{N}$$

The parallel between the term at the right and the expression for a variance should be obvious. A variance is of the form $\Sigma x^2_1/N$ or $\Sigma x^2_2/N$. A covariance is the mean of the cross products of deviations; a variance is a mean of the squares of deviations. With this new information as background, we may translate equation (16.17) into English by saying that the variance of a composite is equal to the sum of variances of the components plus twice the covariances of all pairs of those components. This is a general principle that is important to remember.

From equation (16.17) it follows, by taking square roots, that

$$\sigma_s = \sqrt{\sigma^2_1 + \sigma^2_2 + 2r_{12}\sigma_1\sigma_2} \quad \text{(Standard deviation of the sum of two unweighted measures)} \qquad (16.18)$$

A demonstration of how this works out in a particular sample is given in Table 16.9. Ten scores are given for the same individuals in X_a and in X_b between which the correlation r_{ab} equals zero. If $r = .0$, the third term in formula (16.17) drops out and the variance of the composite is merely the sum of the variances of the components.

In the illustration in Table 16.9, the variances of the two components are 4.2 and 6.6, respectively. Their sum is 10.8, which checks with the mean of the square found from variable X_c. The way in which variances combine is also demonstrated in Fig. 16.3, which pictures hypothetical distributions for X_a, X_b, and their sum X_c. The position of the scale for X_c is determined by the juncture of the lines erected at distances of 1σ from the means of X_a and X_b. The slanted scale of X_c is closer to that of X_b, consistent with the fact that X_b contributes more variance to it than does X_a and the fact that the composite correlates higher with X_b than with X_a. But these are incidental considerations here. The

TABLE 16.9.—THE VARIANCE AND VARIABILITY OF A COMPOSITE SCORE THAT IS THE UNWEIGHTED SUM OF TWO UNCORRELATED SCORES

Individual	X_a	x_a	x^2_a	X_b	x_b	x^2_b	X_c $(X_a + X_b)$	x_c	x^2_c
A	1	−4	16	6	0	0	7	−4	16
B	3	−2	4	7	+1	1	10	−1	1
C	4	−1	1	4	−2	4	8	−3	9
D	5	0	0	10	+4	16	15	+4	16
E	5	0	0	8	+2	4	13	+2	4
F	5	0	0	0	−6	36	5	−6	36
G	5	0	0	6	0	0	11	0	0
H	6	+1	1	8	+2	4	14	+3	9
I	7	+2	4	5	−1	1	12	+1	1
J	9	+4	16	6	0	0	15	+4	16
Σ	50	0	42	60	0	66	110	0	108
M	5.0		4.2	6.0		6.6	11.0		10.8
σ			2.05			2.57			3.29

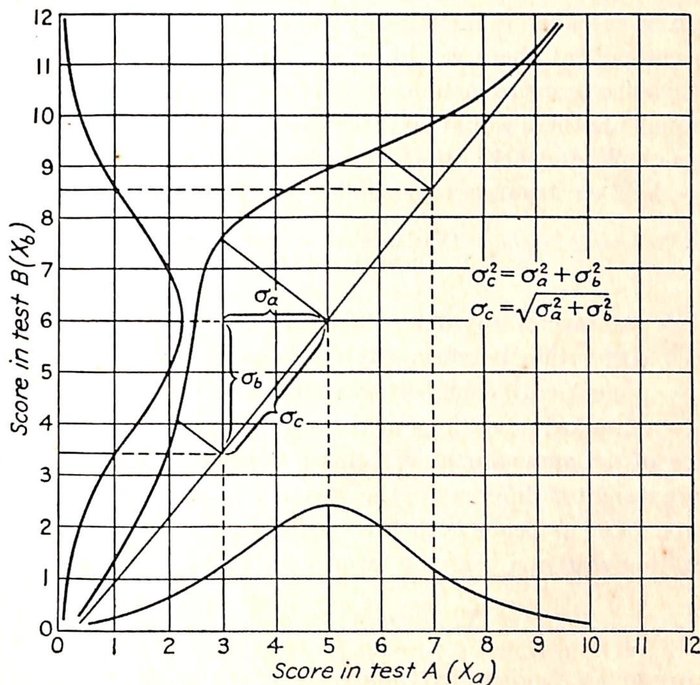


FIG. 16.3.—Illustration of the way in which the standard deviation of an unweighted sum of two scores is related to the standard deviations of those two scores taken separately, when the two are uncorrelated.

important demonstration is that when two variables like X_a and X_b are uncorrelated, we may regard the standard deviation of their composite X_c as the hypotenuse of a right triangle of which σ_a and σ_b are the legs. The old, familiar Pythagorean theorem thus applies to the summation of two independent variables.

Relation of σ_s to the Standard Error of a Difference.—The similarity between equation (16.18) and equation (9.31) for the standard error of a difference will probably have been noticed. The only difference is in the algebraic sign of the covariance term, $2r_{12}\sigma_1\sigma_2$, which is positive in the case of σ_s and negative in the case of σ_d. Of course, in the preceding discussion of σ_s we have been applying it to distributions of single observations, whereas σ_d has been applied to distributions of means (mean differences). The principles are the same, either with means or with single observations. Had we written the summation equation in the form $X_c = X_a - X_b$, instead of $X_c = X_a + X_b$, we would have been dealing with differences instead of sums. On the other hand, in the equation $X_c = X_a - X_b$, we can say that we actually have a summation of scores, those for X_a having a weight of $+1$ and those for X_b a weight of -1.

Variance of a Composite of More than Two Components.—Equation (16.17) can be extended to include any number of unweighted components. For each component there would be its variance but there would be as many covariance terms to include as there are *pairs* of components. With three components there would be three covariance terms: $2r_{12}\sigma_1\sigma_2$, $2r_{13}\sigma_1\sigma_3$, and $2r_{23}\sigma_2\sigma_3$. Where there are n components, there are $n(n-1)/2$ pairs and $n(n-1)/2$ covariances to consider. In terms of a general formula

$$\sigma^2_s = \Sigma\sigma^2_i + 2\Sigma r_{ij}\sigma_i\sigma_j \qquad \text{(Variance of a sum of any number of unweighted components)} \qquad (16.19)$$

where σ^2_i = variance of any one component, X_i.

r_{ij} = correlation between any component X_i and any other component with a higher subscript number.

σ_i and σ_j = standard deviations of the two components correlated.

Variance of a Composite of Weighted Components.—When the components are weighted differently, the variance of the composite will reflect the weights. Let us begin with the special case of two components. If the summation equation is of the form

$$X_{ws} = w_1X_1 + w_2X_2$$

the variance of X_{ws} is given by the equation[1]

[1] For proof, see Appendix A.

$$\sigma^2_{ws} = w^2_1\sigma^2_1 + w^2_2\sigma^2_2 + 2r_{12}w_1\sigma_1w_2\sigma_2 \quad \text{(Variance of a composite of two weighted components)} \quad (16.20)$$

where w_1 and w_2 = weights applied to components X_1 and X_2, respectively.

As an example of this type of problem, let us use the data on X_4 and X_5 in Table 16.1. If these two variables are used in a composite to predict X_1, the least-square solution gives b weights of .224 and .491, respectively, and a multiple R, based upon these weights, of .578. The predicted X values based upon the equation $X'_1 = .224X_4 + .491X_5$ would be expected to have a standard deviation equal to $R_{1.45}$ times σ_1. This product is $.578 \times 9.1$, which equals 5.26. Let us see whether formula (16.20) will lead to the same result. By substituting the appropriate values,

$$\begin{aligned}\sigma^2_{ws} &= (.224^2)(19.4^2) + (.491^2)(3.7^2) + 2(.345)(.224)(19.4)(.491)(3.7)\\ &= (.050176)(376.36) + (.241081)(13.69) + (.690)(4.3456)(1.8167)\\ &= 18.8842 + 3.3304 + 5.4473\\ &= 27.6319\end{aligned}$$

from which

$$\sigma_{ws} = 5.26$$

This agrees exactly with the expectation.

With weights of +1 for both X_4 and X_5, application of formula (16.17) would have given

$$\begin{aligned}\sigma^2_s &= 19.4^2 + 3.7^2 + 2(.345)(19.4)(3.7)\\ &= 439.5782\end{aligned}$$

from which

$$\sigma_s = 21.0$$

Variance of a Composite of Any Number of Weighted Components.—When there are more than two components, each weighted differently, the variance of the composite is given by the general formula[1]

$$\sigma^2_{ws} = \Sigma w^2_i\sigma^2_i + 2\Sigma r_{ij}w_i\sigma_iw_j\sigma_j \quad \text{(Variance of a sum of any number of weighted components)} \quad (16.21)$$

where w_i = weight assigned to variable X_i, where i takes on values 1 to $n - 1$ in turn.

r_{ij} = correlation between X_i and any other variable X_j, where j is a subscript greater than i.

σ_i and σ_j = standard deviations of X_i and X_j, respectively.

[1] See Appendix A.

We could apply formula (16.21) to the four components of the regression equation predicting freshman grades with the appropriate b weight substituted for w in each case. We should find that the standard deviation is equal to R times σ_1, which is $.698 \times 9.1 = 6.35$. The inclusion of variables X_2 and X_3 in the regression equation raises the dispersion of the predicted grades from 5.26, which it would be with X_4 and X_5 only, to 6.35.

Achieving Any Desired Standard Deviation in a Composite.—In using regression equations, the dispersion of the predictions falls short of that of the obtained values. This is all right and proper when we are interested in predicting an individual's most probable measure on the scale of obtained measures in X_1. The regression of predictions toward the general mean is a natural phenomenon of imperfect correlation, as was pointed out before (Ch. 15). There may be other uses of composites, however, that call for other values than those given by the regression equation. Suppose that we wanted predictions to spread just as much as the obtained values do. Suppose that we should want them to be dispersed with some standard variability, for example, with a σ of 10.0, as on a T scale, or a σ of 2.0, as on a C scale. The way that kind of goal can be achieved will now be explained.

Fortunately, for the solution of this problem, it is not the absolute sizes of the weights that matter; it is their ratios to one another. So long as they bear the same relations to each other, the correlation of the composite with some criterion will remain the same. Consequently, we could double, triple, or otherwise change the regression weights by some common multiple, without affecting the predictive value, if all we want is to predict individuals in the same relative positions in a distribution.

The σ of the predictions is always related to the σ of the obtained values by the extent of the correlation (when optimal weights are used). In a multiple-regression problem, σ of the predicted values equals R times the σ of the obtained values. We can therefore make the σ of the predictions equal the σ of the obtained values by dividing each regression coefficient by R. A readjusted b coefficient, then, would be computed by the formula

$$b'_{12.34\ldots m} = \beta_{12.34\ldots m}\left(\frac{\sigma_1}{\sigma_2 R_{1.23\ldots m}}\right) \quad \text{(Regression coefficient adjusted to make the } \sigma \text{ of a composite equal } \sigma_1\text{)} \qquad (16.22)$$

If the σ desired in the composite is 10, or 2, or any other chosen quantity, this could be achieved by substituting that quantity for σ_1 in formula (16.22).

Achieving Any Desired Mean for a Composite.—In the complete regression equation, in order to make the mean of the predictions equal that of the obtained values, the a coefficient is introduced. The computation of

a is given by formula (16.15). After one has determined any weights whatever to apply to the raw scores of the components of a composite measure, the same formula can be applied, putting in the place of M_y any desired quantity. This is true because of the reasoning involved in the computation of the mean of a composite (see formula 16.16). Thus, if we had wanted the mean of the grades predicted by the regression equation on p. 447 to be 50, we would have substituted 50 instead of 73.8, the actual mean of the grades. The only practical restriction would be to choose a mean such that no composite measures would be negative. This means that any chosen mean should be at least 2.5 to 3.0 times the standard deviation of the composite.

Substitutes for Regression Weights.—While regression weights derived from least-square solutions, or weights proportional to them, yield the greatest accuracy of prediction from the variables available, it is often expedient in the practical situation to deviate from the refined solution. It can be shown that we may substitute weights that approximate the regression coefficients, even very roughly at times, and still not affect the degree of correlation very much. Instead of applying weights to three decimal places, one significant digit will often suffice, in other words, simple integral weights. In predicting freshman grades from high-school average and interest score combined, for example, we found the optimal weights to be .224 and .491. We might in practice round these to .2 and .5, respectively. It will be shown later[1] that the change in correlation between X_1' and X_1 in the two cases is from .578, with the three-digit weights, to .577, with the one-digit weights. Surely, this loss is quite trivial. We could use weights of 2 and 5 had we so chosen. Suppose we want even a simpler ratio of the two weights, like 1/2, rather than 2/5. With weights of 1 and 2, also, the correlation of composites and grades would be .577. With equal weights the correlation would drop to .570. Even this much loss could be tolerated.

Before the reader draws the conclusion from this isolated example that all differential weighting is unnecessary, however (many generalizations, unfortunately, are just as sweeping as this would be), it is necessary to consider some points not yet brought out. There is no reason to believe that this is a typical example. Ordinarily, the more independent variables in a composite, the more can one depart from the weights demanded by least-square solutions and yet maintain a high level of correlation between that composite and a criterion to which the weights apply. This is why with a test of many items we may forget to bother with differential weight-

[1] Methods for correlating composites or sums, either weighted or unweighted, will be described beginning on p. 462.

ing. In a two-variable composite, however, we have the minimum number. We would therefore expect to find the validity of the composite to be rather sensitive to changes in weights. Roughly, the explanation in this example is that X_4 (high-school average) has a beta weight about 2.4 times that for X_5 (interest score) and it has a standard deviation about 5 times as large as that for X_5. Even when X_4 and X_5 have the same weight in the composite, X_4 contributes to the composite in proportion to its standard deviation. This follows from equation (16.17) in which it is shown that *without* differential weights *each part's contribution to total variance is proportional to its own variance.* Without differential weighting factors in the equation, then, X_4 is still weighted much more than X_5. This illustrates a fact that is not often realized. It is usually assumed that merely summing several scores weights those scores equally. As a rule, it does not; *it weights them in proportion to their standard deviations.* In more common-sense language, tests weight themselves.

Weighting Measures Inversely as Their Standard Deviations.—This discussion leads to the conclusion that if we really want to weight tests in a battery equally we should apply to each one a weight inversely proportional to its standard deviation. Without information as to the validities of the tests and of their intercorrelations, that would be a reasonable thing to do. It is sometimes done. Table 16.10 shows how this end may be achieved. The four tests are the same as those used to predict freshman grades. The means and standard deviations are duplicates of those given in Table 16.1.

TABLE 16.10.—THE PROCESS OF WEIGHTING COMPONENTS INVERSELY AS THEIR DISPERSIONS

	Variables			
	A	*B*	*C*	*D*
M	19.7	49.5	61.1	29.7
σ	5.2	17.0	19.4	3.7
$19.4/\sigma$ (w')	3.73	1.14	1.00	5.24
Integral weight (W)	4	1	1	5
Estimated importance (I)	2	2	5	1
Combined weight (Iw')	7.46	2.28	5.00	5.24
Revised integral weight (W')	7	2	5	5
Simplified weight ($Iw'/2.28$)	3	1	2	2

We could find a weight equal to $1/\sigma$ for every test, but these weights would be rather small decimal numbers in some cases. A good practical

procedure is to select the largest σ in the list, in this case 19.4, and to compute the ratio $19.4/\sigma$ for each test. The test with the largest σ will have the smallest weight. With this particular ratio, the smallest weight will then be exactly 1.0. The ratio of any other weights to this one will be immediately apparent. It is recommended that all these ratios be rounded to the nearest integer, as shown in the fourth row of Table 16.10. The weights obtained by this process are 4, 1, 1, and 5, respectively. With these weights applied, all four tests would contribute approximately the same amount of variance to the total variance.

The principle of weighting each test inversely as its dispersion is involved in the b coefficient. Remember that b is equal to beta times σ_1/σ_i, where σ_i is the standard deviation of the test to be weighted. Using this procedure, therefore, is virtually equivalent to using an incomplete b coefficient. It virtually assumes equal validities for all tests and equal intercorrelations, conditions which would lead to equal betas.

From the solution in Table 16.10, measures X_4 and X_5 should receive weights of 1 and 5, respectively. The difference is in the same direction as for the two b weights, which are .224 and .491, respectively, but X_4 is given relatively about half as much importance as it should have. The effect upon the correlation of the composite, weighted this way, is to reduce it from the optimal R of .578 to a correlation of .558. The underweighting of X_4, which is more valid and has a larger beta than X_5, shows up in the lower validity of this composite.

Other Principles of Weighting.—Common sense may suggest that component tests should be weighted in proportion to their lengths or their means or other obvious properties. To do so may lead the uninformed investigator astray. If two tests of unequal length are equally effective, in the sense that they produce dispersions in proportion to their lengths, when no weights are applied at all they are automatically weighted in proportion to their lengths. Attaching more weight to the long test thus merely exaggerates an effect we already have. There is no real justification for weighting tests in proportion to their means, and, when means are proportional to standard deviations, the policy would again carry the weighting further in the same direction.

If parts are regarded as *really* of equal importance, then a correction such as was described above would be in order. If the traits measured by different tests are regarded as differing in importance, and if we can decide upon ratios of importance, we can combine weights based upon these ratios with whatever weights we already have. Suppose, for example, we thought that the four variables in Table 16.10 are important in the ratios 2, 2, 5, and 1. Two weights for a variable are combined by finding their

product. In Table 16.10, it would be best to use the factor $19.4/\sigma$ for each test as the weight already established and to multiply it by the weight representing importance. The four products in Table 16.10 are 7.46, 2.28, 5.00, and 5.24, respectively. Rounding these, we have 7, 2, 5, and 5. To simplify these still more, if we let the smallest weight equal 1, the others can be expressed as integral multiples of 1 (found by dividing every product by 2.28). The simplified, combined weights are then 3, 1, 2, and 2. These examples are given merely to illustrate several ways in which weights can be derived to meet different requirements and considerations.

Some investigators believe it important to consider reliabilities of measures in weighting them in combinations. By reliability here is meant consistency of scores as indicated by some kind of a self correlation. If regression weights have been computed, reliabilities have been automatically taken into account and no modification of the weights for reliability would be necessary. But if some other method is used to arrive at weights and if the measures combined differ markedly in reliability, then some index of reliability should be considered. This tends to avoid giving "errors of measurement" in the less reliable instruments too much weight. If reliability coefficients have been computed, the weight contributed from this source should be the square root of each reliability coefficient, rather than the reliability coefficient itself. The type of reliability coefficient should be one indicating internal consistency, *i.e.*, an odd-even type or a Kuder-Richardson type (see Ch. 17).

The Correlation of Composite Measures with Other Measures.—The multiple R is only one index of correlation between a composite measure and some other measure. It applies to a composite in which the weighting has been optimal, with weights determined by the least-square solution. To test the predictive value for composites with other than optimal weights, we have other procedures known under the heading of *correlation of sums*. The components may be unweighted (*i.e.*, each weight is +1) or differentially weighted.

Correlation of a Composite of Unweighted Measures.—The simplest case is solved by the equation[1]

$$r_{cs} = \frac{r_{c1}\sigma_1 + r_{c2}\sigma_2}{\sqrt{\sigma^2_1 + \sigma^2_2 + 2r_{12}\sigma_1\sigma_2}} \qquad (16.23)$$

(Correlation of a sum of two unweighted components with a third variable)

where σ_1 and σ_2 = standard deviations of the two components.

r_{c1} and r_{c2} = correlation of each component with the third variable

[1] See Appendix A for proof.

Let the illustrative summation equation be $X_s = X_4 + X_5$, where X_s stands for a sum of X_4 and X_5, which in recent illustrations have stood for high-school average and interest scores, respectively. What is the correlation of X_s with freshman grades, which here are symbolized by X_c? Applying formula (16.23),

$$r_{cs} = \frac{(.546)(19.4) + (.365)(3.7)}{\sqrt{19.4^2 + 3.7^2 + 2(.345)(19.4)(3.7)}}$$
$$= \frac{11.9429}{\sqrt{429.5782}}$$
$$= .570$$

When there are more than two components, the more general formula for the same kind of correlation is

$$r_{cs} = \frac{\Sigma r_{ci}\sigma_i}{\sqrt{\Sigma\sigma^2_i + 2\Sigma r_{ij}\sigma_i\sigma_j}} \qquad (16.24)$$

(Correlation between a sum of unweighted variables and another single variable)

where r_{ci} = correlation between any one component X_i and the outside single variable (i varies from 1 to n).
σ_i = standard deviation of the same component.
r_{ij} = correlation between X_i and any other component X_j, when j is a higher subscript number than i.[1]

Correlation of a Composite of Weighted Measures.—When there are two components, each weighted differently, the correlation with a third measure is given by[2]

$$r_{c(ws)} = \frac{w_1 r_{c1}\sigma_1 + w_2 r_{c2}\sigma_2}{\sqrt{w^2_1\sigma^2_1 + w^2_2\sigma^2_2 + 2r_{12}w_1\sigma_1 w_2\sigma_2}} \qquad (16.25)$$

(Correlation of a sum of two weighted measures with a third measure)

where w_1 and w_2 = weights attached to measures X_1 and X_2, respectively, and other symbols are as defined in formula (16.23).

For the combination of high-school average and interest scores, let us assume weights of 2 and 5, respectively. These are closely proportional to the b coefficients of .224 and .491, respectively. Applying formula (16.25),

$$r_{c(ws)} = \frac{2(.546)(19.4) + 5(.365)(3.7)}{\sqrt{4(19.4^2) + 25(3.7^2) + 2(.345)(2)(19.4)(5)(3.7)}}$$
$$= \frac{27.9373}{\sqrt{2342.972}}$$
$$= .577$$

[1] Here, as in similar formulas, $r_{ij}\sigma_i\sigma_j$ implies covariances of all possible pairs of variables.
[2] For proof, see Appendix A.

Thus, crude, integral weights of 2 and 5 would give as high a correlation of the combination of X_4 and X_5 with X_1 (freshman grades) as would the three-digit *b* coefficients .224 and .491.

For the general case, with more than two components, the correlation with an outside variable is

$$r_{c(ws)} = \frac{\Sigma w_i r_{ci} \sigma_i}{\sqrt{\Sigma w^2{}_i \sigma^2{}_i + 2\Sigma r_{ij} w_i \sigma_i w_j \sigma_j}} \quad \text{(Correlation of a weighted sum with an outside variable)} \qquad (16.26)$$

where the symbols are as defined in preceding formulas.

Alternative Summarizing Methods

Summative equations represent only one way in which several measures may be combined in order to reach single predictions or decisions. There are alternative methods some of which are better then regression equations in certain situations. The two chief contenders are the multiple cutoff method and the profile method. These will be described and their variations discussed.

Multiple-cutoff Methods.—In a multiple-cutoff method, a minimum qualifying score or measure is adopted for each variable used in making a joint prediction. A good example of the method is the medical examination in the qualification of individuals for military service, for life insurance, or for employment. Failure to meet the standard on any one test may disqualify the individual. Making a particularly good showing in one respect is not ordinarily allowed to compensate for a poor showing in some other. The phenomenon of compensation, which the regression-equation approach allows, is the chief difference between the two methods, in principle.

Multiple Cutoffs Contrasted with Multiple Regression.—A geometric illustration of the difference between the two methods may be seen in Fig. 16.4. The two variables represented there (X_2 and X_3) are both independent variables, used jointly to predict some criterion X_1 which is not shown. A moderate correlation, of approximately .40, is assumed between X_2 and X_3, as represented by the familiar elliptical distribution of the population. Let us assume a selection problem and that we have the alternative of applying two cutoff scores X_{2_c} and X_{3_c} or of applying a single cutoff score based upon a weighted sum of X_2 and X_3. Assume also that we reject the same proportion of the applicants by either method.

The use of two cutoff scores would reject all individuals to the left of the point X_{2_c} and a vertical line erected at that point, also all individuals below the point X_{3_c} and a horizontal line drawn at that level. Some indi-

viduals would be rejected on the basis of either variable alone and some on the basis of failure to meet standards on both. The single cutoff on the weighted composite, however, would be represented by a slanted line. This is consistent with the slanted-line system shown in Fig. 16.2. All individuals below and to the left of this slanted line would be rejected.

It is now possible to see what kind of individuals would be accepted by the one method and rejected by the other and on which ones the two methods agree. The individuals in area A of the ellipse would be accepted by either method. The individuals in area R would be rejected by either

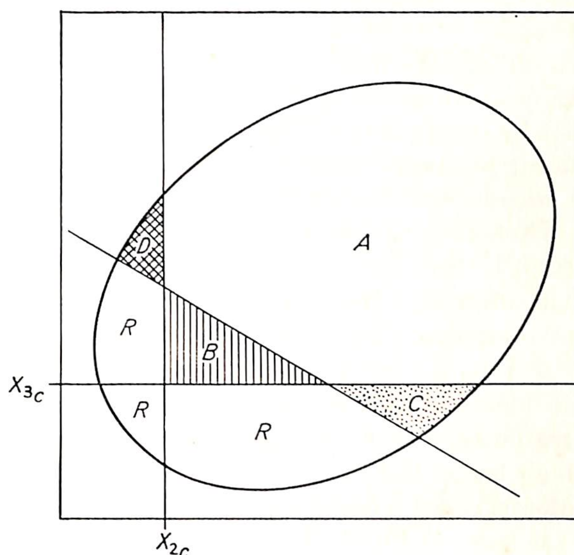


FIG. 16.4.—Geometric comparison of accepted and rejected personnel by the multiple-regression-equation method and by the multiple-cutoff method, when approximately equal proportions are selected by either method. (*After R. L. Thorndike, AAF Report No.* 3.)

method. Individuals in area B would have been rejected by the multiple-regression-equation method but would be accepted by the multiple-cutoff method. Individuals in areas C and D would be accepted by the regression method but rejected by the cutoff method, those in C for different reasons than those in D.

The crux of the comparison of values of the two methods lies in determining whether individuals in area B are any better in the criterion than those in areas C and D. Individuals in area B are rejected by the one method because they combine below-average scores in X_2 and X_3. They just succeed in meeting minimum standards in both variables and so would be accepted by the other method. Individuals in areas C and D,

although below standards in one variable, are allowed to present compensating strong scores in the other variable and hence to be accepted by the one method. They are regarded as doubtful risks by the other method.

It can be argued that not enough is known about compensatory effects in performances that serve as criteria, and that is quite true. There should be some experimental studies of this kind. A vindication of the regression method, however, is found in the consistency with which composite scores continue to correlate as they do in line with multiple-correlation coefficients that forecast those correlations. If compensatory effects did not occur, there would probably be much more shrinkage in correlation of sums with criteria than there is.

An Evaluation of the Multiple-Cutoff Method.—If all regressions are linear, theoretically, there should be no advantage in selection by multiple cutoffs over that by composites. This can be explained roughly by the fact that in a linear regression there is a *continuous* improvement in criterion measures with increased score in an independent variable, and at a constant rate. Thus, so far as the relationship between the test and the criterion is concerned, there is no more reason for putting the cutoff at one point rather than another. The cutoff would have to be established on the basis of some other determiners, such as success ratio or validity. In using a number of tests for selection for a single purpose, presumably it would be best to make the most rejections on the basis of the most valid test. When a regression is definitely curved, there is a real basis for using a cutoff on a single test. The cutoff would be established in line with the region of transition between low and high rates of increase in the criterion measure. For example, in Fig. 15.12, somewhere between the scores 90 and 100 would be a good division point, taking advantage of the rapid increase in criterion values as scores increase in X, and at the same time recognizing that above a score of 100 there are no appreciable differences in criterion values as X changes.

There are some practical difficulties in the administration of multiple cutoffs which make the method less appealing than a regression equation. There is the difficulty of establishing several different cutoff points which will take full advantage of the differences in validity among the tests and which will yield the appropriate numbers of qualified applicants. Once the minimum standards are established, however, the method is simple to apply. Failure to meet any one of the minimal scores automatically means rejection.

Rejection of an applicant on the basis of a single test is somewhat risky as compared with rejection on the basis of a composite score because of the fact that the reliability of a single test score is usually less than that for a

composite. If the parts of a composite are positively intercorrelated, the total score is more reliable than the part scores.

Some Variations of the Multiple-cutoff Method.—A distinction is made between a *simultaneous-hurdles* method and a *successive-hurdles* procedure in testing programs using multiple cutoffs.[1] In the former, all applicants take all tests; in the latter they do not—they continue to take tests only as long as they continue to qualify on them. After the first failure they are rejected. In the latter method it is good practice to administer the most valid test first. It is the one on which the largest number of rejections should be made. It is desirable, too, that if a single attempt is to be decisive for so many individuals, the decision should be made on as good a basis as possible. If a test of very low validity were given first, some who could qualify on the valid test would never have a chance to take it. Such individuals might be expected to fail when they took the invalid test later, of course, but remember that tests are not perfectly reliable, and a person might pass a certain test on one day and fail it on another. The successive-hurdles method has the great practical advantage of saving in testing time. If there are many more applicants than openings, large numbers of applicants can be screened and eliminated from further testing by means of a single preliminary examination.

Other variations in using the multiple-cutoff principle have to do with rules concerning rejection. It is not necessary to base a rejection on one test alone. The rules might allow for failure on not more than two, or any selected number of variables. The rules might be refined to the extent of considering pairs or triads of tests. Rejection might be reserved for those who fail on test *M* only if they also fail on test *N*, and so on. Such refinement, however, must be based upon good evidence that it pays in terms of better selection. For most purposes such evidence is lacking.

Profile Methods.—For guidance work and clinical work in general there is common preference for seeing an individual's scores represented in a pattern provided in a profile. A single summative score is unsuitable or may be unobtainable. A single composite score is unsuitable perhaps because the problem is not a selection problem but a classification problem. In vocational guidance, clients are "sorted" into vocational categories. If there were single summative scores already established with satisfactory correlations with vocational criteria of many kinds, perhaps the profile method would be less important. Clinicians commonly express a desire to "see a personality in its totality," however, and a profile is one approach to this end.

[1] Toops, H. A. Philosophy and practice of personnel selection. *Educ. & Psychol. Meas.*, 1945, **5**, 95–124.

There are several ways of using profiles. Some prefer the intuition given by a general impression of a plotted graph for an individual. Others prefer to match more definitely described job-requirement or adjustment-requirement patterns with individual trait patterns. It is possible, by means of careful research, to define certain adjustment requirements in terms of optimal scores in a number of different variables. This statement implies curved regressions, and that is precisely the condition which favors the choice of a profile method to a regression method.

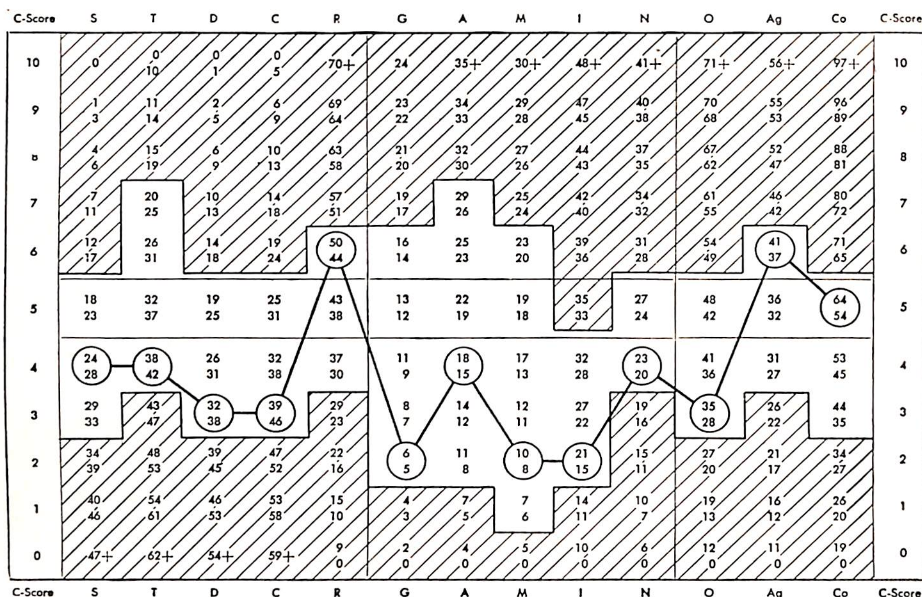


FIG. 16.5.—An illustration of the profile method of selection applied to personality-inventory scores. The clear portion of the chart represents what is believed on the basis of experience to be the most favorable score ranges for personnel who are assigned to a certain routine type of work. The scores of the worker shown all fell within the favorable region. (*Courtesy of R. P. Kreuter, Hand Knit Hosiery Company, Sheboygan, Wisconsin.*)

Figure 16.5 demonstrates this kind of use of a profile. By experience, it was found that female workers in a certain kind of routine task tended to be most suited to the job if they had scores in certain regions on the 13 traits scored in the Guilford-Martin personality inventories. Such workers were likely to be best if somewhat shy or reclusive, a little on the depressed and emotional side, less active than average (the task was sedentary), less ascendant socially, somewhat beset with feelings of inferiority, somewhat subjective or hypersensitive, and perhaps none too agreeable or cooperative. In most respects the tendencies listed would seem to picture a generally "poor" personality picture. Low extremes were unfavorable, however; the general tendency was just average or slightly below in most

traits. This is understandable in that such an individual is probably lacking in aspirations for positions that require the better qualities and is contented with a routine type of work in which adjustments to social requirements are relatively easy. The profile is shown of a certain individual who was rated very high in performance at her task.

For selection purposes, a profile may be handled in various ways. The one shown in Fig. 16.5 illustrates one procedure. The favorable zone is clear, and less favorable zones are crosshatched. The crosshatching can be overprinted on the chart or a plastic mask can be prepared to lay over individual charts. Decisions can be based upon the *number* of favorable scores or upon the trend of the individual's curve as compared with the trend of the optimal scores. If a single optimal score has been determined for every trait, and an "ideal" profile has been drawn, the departure of a single profile from the ideal profile can be determined in various ways, none of them highly satisfactory. The deviations of each person's scores from the ideal scores can be summarized in various ways. A way that meets common statistical principles would be to square the deviation, sum the squares, find a mean, and then a square root. This would give a single summarizing statistic that has some statistical sanction. There are many who would want more than such a number, however, for it does not tell us where the deviations are. The general problem of using profiles in a rigorous manner is still unsolved.

Classification of Personnel.—Selection of personnel presupposes a supply of applicants and the possibility of rejecting a proportion of them. Attention is upon one kind of assignment to be filled. In the classification of personnel, there are two or more assignments that can be made and one might even consider rejecting none, provided proper assignments can be found for all. In some situations there is the double problem of selection and classification combined. The availability of more than one assignment, however, makes possible the utilization of many more applicants than would be true if there were only one kind of place to fill, for, presumably, personnel who do not qualify for one place might well qualify for some other. The more different kinds of places there are to fill, the smaller the chance of any applicant's being rejected for every kind.

Classification, broadly defined, means assigning individuals each to his appropriate category. This would include the operations in educational and vocational guidance. In vocational guidance, the number of kinds of "assignments" is almost infinite, though the number of major categories is limited. In selection we have an assignment with the need to find the person for it; in classification in general, we have a number of assignments with their requirements in terms of human resources, on the one hand, and

a number of persons who have the resources to satisfy or not to satisfy each assignment on the other. In vocational guidance, we have one individual, with a unique pattern of resources, on the one hand, and a large variety of possible occupations or assignments, on the other.

As demonstrated in this and in preceding chapters, we have solved many of the statistical problems involved in selection. These are bound up with the problems of prediction and of how to evaluate the goodness of a prediction. By contrast, the problems of classification have been seriously neglected. Assignment to alternative classes requires a *differential prediction* rather than a prediction of a single variable. We have to predict how much better the individual will adjust or perform if assigned to one category than if assigned to some other category. There are no regression equations as yet devised for this particular purpose, either from a single pair of single variables or from two composites. Presumably, when only two assignments are being considered and two predictive indices, we attempt to predict a *difference* in the criterion variable from a *difference* in the assessment variable. It is reasonable that the more independence there is between two criterion variables the more easily one could make a differential prediction and the more confidently would one expect to find relatively independent assessment variables. Lack of correlation between assessment variables would seem to be just as important as independence between criterion variables. This is probably not the whole story, as a preliminary study of this problem by R. L. Thorndike has shown.[1]

The problem does not seem so very difficult when only two qualitative categories of classification are involved. When the number of categories increases, the number of pairs becomes rapidly enormous and the differential predictive problem becomes highly complicated. Profiles may represent one answer. This would mean the standardization of profile patterns for different assignments. Composite scores is another kind of approximate solution. From a large battery of tests we can derive a unique composite score for prediction of adjustment or of success in each category. The use of such scores means essentially a classification indirectly through selection. That is, the assignment of each person would be made with reference to his highest composite scores which should be taken into consideration with other information. If there were a sufficient number of composites, each person would probably be found fitted for at least one assignment.

[1] Thorndike, R. L. Research problems and techniques. AAF aviation psychology research program, Report No. 3. Washington, D.C.: Government Printing Office, 1947, 125.

Exercises

1. Using Data 16*A*, compute a regression equation involving X_1, X_2, and X_4. Present beta coefficients, multiple *R*, and other necessary statistics. Interpret your results.

2. Using Data 16*B*, compute a regression equation involving X_1, X_3, and X_5. Present all statistics as computed in an ordinary solution to a multiple-correlation problem. Interpret your results.

3. Give Data 16*A* a complete solution, using the Doolittle method. Include a regression equation and your interpretations.

4. Do the same for Data 16*B* as was called for in Exercise 3.

5. Find the best combination of three predictive indices for either Data 16*A* or Data 16*B*.

6. For Data 16*A* or Data 16*B*, assume five reasonable sets of scores for five hypothetical individuals in the independent variables for which you have solved the regression equation, and from them predict X_1 values.

7. Compute standard errors of multiple estimate, coefficients of multiple determination and nondetermination, and indices of forecasting efficiency for the problem in Exercises 1 and 3. Interpret your results.

8. Determine how much of the total variance in the dependent variable is accounted for in the composites in Exercises 2 and 4. Interpret your results.

9. Compute standard errors of the regression coefficients in Exercises 1 and 2. Draw conclusions.

10. Apply shrinkage formulas to the results in Exercises 3 and 4, both in the multiple *R*'s and in the standard errors of estimate. State conclusions.

DATA 16*A*.—INTERCORRELATIONS OF SCORES FROM FOUR EXAMINATIONS AND MARKS RECEIVED IN FRESHMAN MATHEMATICS
($N = 100$)

Variable	X_2	X_3	X_4	X_5	X_1
X_2	—	.70	.53	.39	.51
X_3	.70	—	.61	.29	.51
X_4	.53	.61	—	.28	.61
X_5	.39	.29	.28	—	.39
X_1	.51	.51	.61	.39	
M_x	4.10	5.44	5.37	4.95	5.70
σ_x	1.92	1.84	2.26	2.14	2.42

X_2 = Ohio State psychological examination.
X_3 = English-usage examination.
X_4 = algebra examination.
X_5 = engineering aptitude examination.
X_1 = marks in freshman mathematics.

11. By the iterative method, solve for the optimal beta weights for either Data 16*A* or 16*B*. Compare them with those found by the Doolittle method.

DATA 16*B*.—INTERCORRELATIONS OF SCORES FROM FOUR EXAMINATIONS AND MARKS RECEIVED IN ENGINEERING DRAWING

(N = 154)

Variable	X_2	X_3	X_4	X_5	X_1
X_2	—	.53	.24	.28	.33
X_3	.53	—	.24	.11	.34
X_4	.24	.24	—	.38	.31
X_5	.28	.11	.38	—	.41
X_1	.33	.34	.31	.41	
M_x	4.19	5.42	4.70	4.85	5.25
σ_x	2.04	2.32	1.93	2.05	1.45

X_2 = Ohio State psychological examination.
X_3 = algebra examination.
X_4 = paper-folding test.
X_5 = form-perception test.
X_1 = term mark in engineering drawing.

12. Estimate the means of the combinations of scores by the regression equations found in Exercises 1 and 2 omitting the constant a from each equation. Do the same for the combinations of scores by equations in Exercises 3 and 4, also omitting the constant a.

13. Estimate the standard deviation of an unweighted sum of scores X_2 and X_4 in Data 16*A*. Estimate the standard deviation of a weighted combination of scores in X_2 and X_4 in the same data, using the optimal weights, in two ways. Use weights of 2 and 5 for the same two variables, respectively, and estimate the standard deviation of such a composite.

14. Find the correlation of an unweighted combination of X_2 and X_4 with X_1, also for a combination with weights of 2 and 5, respectively. Compare these with the multiple $R_{1.24}$.

CHAPTER 17

RELIABILITY OF MEASUREMENTS

The Importance of Reliability.—Much of what was said in previous chapters assumed that measurements were perfectly reliable, or nearly so. By a perfectly reliable measurement we mean one that is completely stable or fixed. The same "yardstick" applied to the same individual or object should yield the same value from moment to moment, provided the thing measured has itself not changed in the meantime. An unreliable yardstick is a "rubbery" yardstick.

There are times, both in theoretical investigations and in practical work, when it is very important to take into account the question of reliability. Although numbers, as such, are exact descriptions, just because we amass a series of numbers attached to individuals or to observations is no assurance that those numbers mean much at all about the things measured. There is no way of just looking at numbers and telling whether or not they stand for any real values or could have been "pulled out of a hat." Some samples of measurements actually approach the chance condition just implied. Others are not exactly "chance" collections of numbers, but there is a strong element of chance involved in them. Conclusions to be derived from the very same statistical results might differ considerably whether we know the measurements to be highly reliable or not. Tests of differences and correlation coefficients may often prove to be insignificant merely because the measures used were lacking in reliability. Thus, the matter of reliability well merits considerable attention.

Reliability Theory

It is impossible to appreciate the many problems that arise in connection with reliability and the several meanings of the term itself without going into some of the mathematical ideas underlying the concept. The reader will find that on the one hand there is a rigorously defined conception of reliability from which it is possible to understand many of the peculiarities of measurements, particularly those called test scores, and on the other hand there are several operational conceptions of reliability, depending upon how it is estimated from empirical data—such as internal-consistency, test-retest, and alternate-forms methods. Keeping in mind

the fact that there are several kinds of reliability and that operational definitions and logical definitions do not coincide will aid a great deal in thinking about problems of reliability. We will begin with the basic, theoretical conceptions of reliability.

The Basic Definition of Reliability.—The reliability of any set of measurements is logically defined as the proportion of their variance that is *true* variance. Before elaborating upon the heart of this statement, which is the last part, attention should be called to the more incidental part. The statement begins with "the reliability of any set of measurements." Note that it is the *measurements* that are said to have the property of reliability rather than the measuring instrument. That is because in psychological and educational measurement, and other social measurements, reliability depends upon the population measured as well as upon the measuring instrument. It can rarely be said of any instrument, test or other device, that *the* reliability of that device is of a certain value (usually in the form of a coefficient of correlation.) *Reliability is of a certain instrument applied to a certain population.*

The next comment on the definition, and a more important one, is in definition of *true* variance. The idea of variance itself is not new. The total variance, which we will now call σ^2_t, of a set of measurements is the mean of the sum of squares of deviations from the mean of the measurements. The idea of separating total variance into components is also not new. That idea was emphasized in the chapter on analysis of variance (Ch. 10) and in the chapters on prediction of measurements (15 and 16). Here we make a new kind of segregation of variances. We think of the total variance of a set of measures as being made up of two sources or kinds of variance: *true* variance and *error* variance. We think of each single measurement, also, as having two contributions to it: a true measure and an error. In terms of an equation,

$$X_t = X_\infty + X_e \quad \text{(An obtained measure expressed as the sum of a true and an error component)} \qquad (17.1)$$

where X_t = obtained score or measure.

X_∞ = true score or measure.

X_e = an error increment or component.

Several assumptions are made in connection with this equation. The *true* measure is assumed to be the genuine value of the thing measured and a value we would obtain if we had a perfect instrument. Another conception is that it is the mean value we would obtain for the object if we measured it a very large number of times. There is no inconsistency between these two conceptions. Any obtained measurement at a particular

moment is determined in part by the true value and in part by conditions which bring about a departure, perhaps, from that value. In measuring a series of objects it is assumed that the error components occur independently and at random, that their mean is zero (they increase as often as they decrease a measurement), and that they are uncorrelated with the true values and with errors in other measurements. The assumption that the mean of the errors is zero is not essential but it is convenient. These conditions may not always be satisfied. Without evidence to the contrary we assume that they are satisfied. Knowledge of the instrument and of the other conditions of measurement is sometimes sufficient to lend support to these assumptions or to cause us to reject them in any particular situation.

Reliability was defined as the portion of the total variance that is true variance. The three variances, true, error, and total, are illustrated in Table 17.1. There we have a set of 10 hypothetical true measures whose

TABLE 17.1.—DISPERSION OF TRUE MEASURES, ERROR COMPONENTS, AND THEIR SUMS, THE TOTAL MEASURES, WITH MEANS, VARIANCES, AND STANDARD DEVIATIONS

	True measures X_∞	Error components X_e	Total measures X_t $(X_\infty + X_e)$
	5	− 2	3
	15	+ 2	17
	20	− 4	16
	25	− 2	23
	25	+ 2	27
	25	0	25
	25	+10	35
	30	− 4	26
	35	− 2	33
	45	0	45
Σ	250	0	250
M	25.0	0.0	25.0
Σx^2	1050	152	1202
σ^2	105.0	15.2	120.2
σ	10.2	3.9	11.0
	σ_∞	σ_e	σ_t

mean is 25 and whose variance is 105.0. For each true measure we have a corresponding error component that is to be added to it to form a total, or

obtained, measure for the individual. The mean of these error components is zero, as assumed above. Their variance is equal to 15.2.

The variance of the total measures can be estimated from the component variances by using formula (16.17) of the preceding chapter. It is merely the sum of the two component variances. In the new symbols,

$$\sigma^2_t = \sigma^2_\infty + \sigma^2_e \quad \text{(A total variance as the sum of true and error variances)} \qquad (17.2)$$

The application of this equation in Table 17.1 gives a total variance of 120.2, which checks with that computed from the sum of squares of X_t.

In satisfaction of the definition of reliability, we need to find the proportion of total variance that is true variance. If we divide equation (17.2) through by σ^2_t, we have proportions:

$$\frac{\sigma^2_t}{\sigma^2_t} = \frac{\sigma^2_\infty}{\sigma^2_t} + \frac{\sigma^2_e}{\sigma^2_t} = 1.00 \quad \text{(Sum of proportions of true and error variance)} \qquad (17.3)$$

In symbolic form, the reliability of these measurements is given by the ratio $\sigma^2_\infty/\sigma^2_t$, or in another form by $1 - \sigma^2_e/\sigma^2_t$. In other words, the reliability is measured by the ratio of true variance to total variance, or by one minus the ratio of error variance to total variance. Letting r_{tt} stand for the coefficient of reliability, we have two alternative equations:

$$r_{tt} = \frac{\sigma^2_\infty}{\sigma^2_t}$$

and (Basic equations for the coefficient of reliability) (17.4)

$$r_{tt} = 1 - \frac{\sigma^2_e}{\sigma^2_t}$$

For the problem of Table 17.1,

$$r_{tt} = \frac{105.0}{120.2} = .87$$

or

$$r_{tt} = 1 - \frac{15.2}{120.2} = 1 - .13 = .87$$

If we let e^2 stand for the proportion of error variance in the total, we have the summational equation

$$r_{tt} + e^2 = 1.00 \quad \text{(Complementary nature of true and error variance)} \qquad (17.5)$$

The previous relationships are demonstrated pictorially in Fig. 17.1 and Fig. 17.2. In the first of these two, dispersions of true measures and of total measures are shown. Both have the same mean. The standard deviation σ_t is greater than σ_∞. This is always true, unless they happen to be equal. The effect of errors of measurement is always to increase obtained dispersions; never to decrease them, unless they should be correlated with the true measures or with each other. This suggests that standard errors of means and other statistics, which are estimated from obtained σ's, are inflated values when measures are at all unreliable. Tests of significance are therefore reduced in power by unreliability. The only remedy is to improve reliability of measures or to increase the size of sample to compensate for errors of measurement. There are no known corrections to apply, nor could they probably be justified.

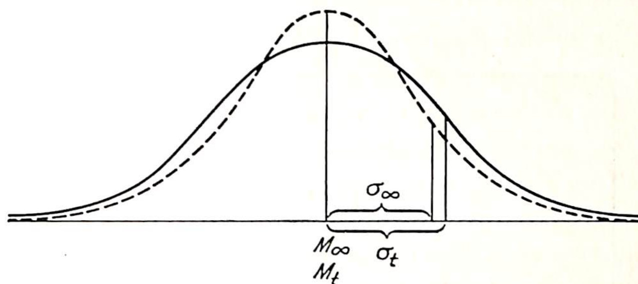


FIG. 17.1.—Distribution of obtained scores in a test (solid curve) and of the hypothetical true components of those scores (dotted curve). Means of obtained and true scores coincide, on the assumption that errors of measurement have a mean of zero. The standard deviation of the obtained scores is larger than that of their true components.

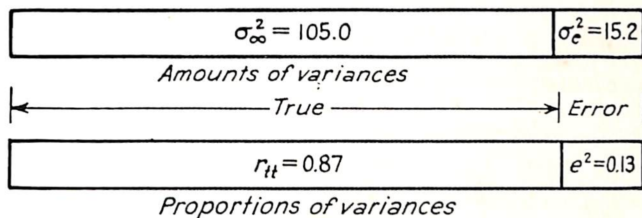


FIG. 17.2.—Amounts of true and error variance (first bar) in a test; also proportions of true and error variance (second bar).

Figure 17.2 presents the picture in a somewhat different manner. Here the summative properties of variances is apparent. Without the assumption of zero correlations for the errors, such a simple picture would be impossible. This kind of representation of variances, in tests particularly, will be encountered with increasing frequency in this and the next chapter.

The Index of Reliability.—The reliability coefficient for a test, r_{tt}, as described thus far, is merely an abstract idea. Operationally, it is some kind of self-correlation of a test, as most textbooks indicate. Before we go into the various operations for estimating r_{tt}, let us add more fundamental meaning to the idea of reliability. Let us think of the true score (X_∞) and the obtained score (X_t) as being two separate variables, the one dependent upon or predictable from the other. This is in spite of the fact that the one includes the other. Think of X_t as the dependent variable

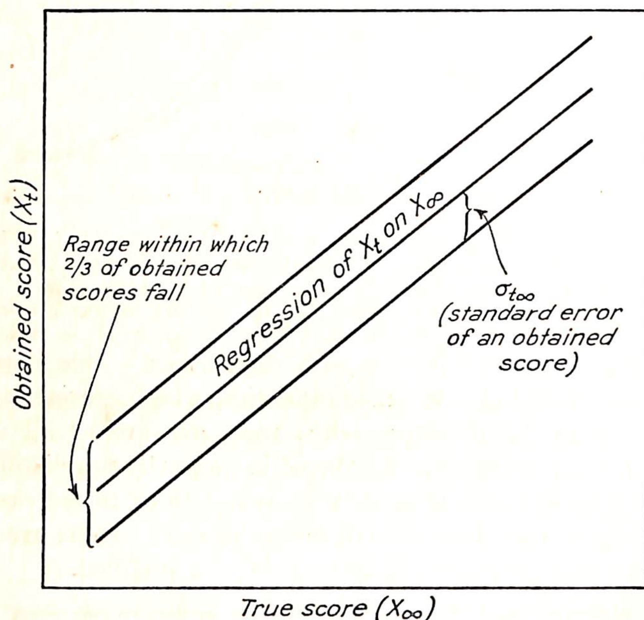


FIG. 17.3.—Regression of obtained scores on true scores, with parallel lines drawn at vertical distances of one standard error ($\sigma_{t\infty}$) from the regression line. (Compare with Fig. 15.6; the standard error of an obtained score is essentially a standard error of estimate.)

and of X_∞ as the independent variable. In a real sense X_t is determined by or dependent upon X_∞. Figure 17.3 shows these two variables as X and Y coordinates and the line of regression of X_t upon X_∞. The correlation between the two is $r_{t\infty}$. The square of this correlation coefficient is an index of determination (see Ch. 15) and it indicates the proportion of variance in X_t that is determined by variance in X_∞. But this is precisely what the reliability coefficient (r_{tt}) tells us. Consequently, we have shown that

$$r^2_{t\infty} = r_{tt} \qquad (17.6)$$

and

(Relation of an index of reliability to a coefficient of reliability)

$$r_{t\infty} = \sqrt{r_{tt}} \qquad (17.7)$$

The correlation between test scores and what they actually measure ($r_{t\infty}$) is called the *index of reliability*. Nothing can correlate with obtained scores higher than their correlation with corresponding true scores. The statistic $r_{t\infty}$, then, is often used as an indication of the higher limit of correlation of any variable with another. Since $r_{t\infty}$ is the square root of the reliability coefficient, it is always numerically higher than r_{tt}. Do not be surprised, then, to find that a test may correlate higher with another than it correlates with itself. We cannot compute $r_{t\infty}$ directly from data, but it can be estimated from r_{tt} or from other information. It is a seldom used statistic, but has a definite meaning and could be used along with r_{tt} or in place of it.

The Standard Error of an Obtained Score.—Since we can estimate the correlation between obtained and true scores and can think in terms of prediction of one from the other, we can also ask concerning the errors of prediction. We know the obtained scores and from them could predict true scores (assuming any mean and standard deviation we please for the true-score scale). But there is nothing to be gained by so doing, for the predictions would be no more accurate than the scores from which they were obtained, and nothing would have happened except a change of unit and zero point.

Suppose that we think in terms of prediction in the other direction; from true scores to obtained scores. This is impossible, since we do not know the true scores from which to make predictions. Let us think rather in terms of determination; of true scores *determining* obtained scores. But errors of measurement also help to determine obtained scores. We are interested in the extent of the discrepancies caused by these errors of measurement, in other words, in the size of distortions produced in the otherwise true-determined measurements. The average of these discrepancies is estimated by the formula

$$\sigma_{t\infty} = \sigma_t \sqrt{1 - r_{tt}} \quad \text{(Standard error of an obtained measure)} \quad (17.8)$$

where σ_t = standard deviation of the distribution of obtained scores.
r_{tt} = reliability coefficient.

The standard error of an obtained score is a standard error of estimate and may be interpreted as such.[1] Figure 17.3 shows the limits marked off at distances of plus and minus one $\sigma_{t\infty}$ from the regression line. In a certain test with a $\sigma_{t\infty}$ equal to 2.0 units, we may say that two-thirds of the obtained scores are within 2.0 units of the true scores that determined them. If a certain individual's true score were 35, for example, the odds are 2 to 1

[1] This statistic is frequently called the *standard error of measurement*.

that his obtained score would not exceed 37 or fall below 33. Allowing a margin of 2σ, we can say that the odds are 19 to 1 that his obtained score will not exceed 39 or fall below 31. Any obtained score does not tell us what the corresponding true score is, but with knowledge of the σ_{t_∞} we can have a degree of confidence that the true score cannot be very far away. The same standard error gives us some basis for confidence as to whether the scores for two persons represent a real difference or whether we can tolerate the idea that they could have come from the same true score.

Reliability at Different Parts of the Test Scale.—Test users frequently ask to know the standard error of an obtained score rather than the reliability coefficient, because it tells them more directly what they wish to know. It tells them whether they should be concerned about differences of 2, 4, 8, or 12 points or whether any or all of these differences are within the probable range that could have been produced by errors of measurement. It may happen, however, that because of a peculiarity of the test itself, discriminations are better at one part of the scale than at other parts. The σ_{t_∞} statistic is a blanket index, implying approximately equal discriminating power all along the scale. If there is reason to suspect that discrimination is actually unequal along the scale, this can be examined by preparing a scatter diagram, showing the relationship between two forms (or halves) of the same test. The standard deviations of the columns or rows at different score levels will indicate where predictions have the greatest accuracy.

Computing the Standard Error of an Obtained Score from Differences.—Rulon has devised a way of computing σ_{t_∞} directly from differences between scores made by individuals on odd and even pools of items.[1] The equation is

$$\sigma_{t_\infty} = \sqrt{\frac{\Sigma d^2}{N}} \quad \text{(Standard error of an obtained measure computed from differences)} \qquad (17.9)$$

where d = difference between two scores of half-tests for one individual. A rough rationale for the Rulon method is to say that a difference between one half-score and the other half-score for the same person is a measure of the error for that individual. Since errors are conceived as deviations, squaring, summing, and dividing by N should estimate the amount of error variance. That is precisely what $\sigma^2_{t_\infty}$ signifies—the amount of error variance. Thus, $\sigma^2_{t_\infty} = \sigma^2_e = \sigma^2_t - \sigma^2_\infty$. This fact will be used later as another way of estimating the reliability coefficient.

[1] Rulon, P. J. A simplified procedure for determining the reliability of a test by split-halves, *Harv. educ. Rev.*, 1939, **9**, 99–103.

Methods of Estimating the Reliability Coefficient

We leave theory for a while and see how r_{tt} can be estimated from empirical data. There are many procedures, falling roughly into the three categories (1) internal-consistency reliability, or simply internal consistency; (2) alternate-forms reliability, or comparable-forms reliability; and (3) retest reliability, or test-retest reliability. Cronbach has recently proposed that we speak of the second and third types of estimate as coefficients of equivalence and of stability, respectively.[1] It would be convenient, also, to speak of the first type as a coefficient of consistency.

There is no one best way of estimating r_{tt}. The type preferred will depend upon one's purposes and the meaning and use one wishes to attach to r_{tt}. A secondary consideration is availability of data in the proper form. Other considerations have to do with testing conditions and the kind of test or other measure.

The nature of procedures differs most in the kind of things that are allowed to be considered as true variance and as error variance. What may be regarded as true variance in computing one kind of r_{tt} may be regarded as error variance in computing one of the others. For the sake of clear thinking, it will pay us to look at some examples of this.

Contributors to True and Error Variance.—On the whole, things that contribute to an examinee making the same score in "repeated" applications of a test are contributors to true variance in the obtained scores. The word "repeated" is in quotation marks because the repetition is broadly defined to include alternate forms or two halves of the same test. On the whole, things that contribute to different evaluations of performance of an individual in a test are contributors to error variance. The sources of either kind of variance are numerous. Certain of them are of sufficient clarity and commonness of appearance to be recognized and named.

Let the bar diagram in Fig. 17.4 represent the total variance in obtained scores of a test. Let c^2 be that proportion of the total variance that would be regarded as true variance no matter what method of estimating r_{tt} is employed. After all, they should have very much in common. Let e^2_a be regarded as those sources of error variance that are unique to the alternate-forms method but are regarded as sources of true variance for the other methods. The relative sizes of these portions will vary from test to test. Actual examples of e^2_a and of c^2 will be given shortly. Let e^2_i be sources of error variance particularly when some internal-consistency

[1] Cronbach, L. J. Test "reliability": its meaning and determination. *Psychom.*, 1947, **12,** 1–16.

method is used. This portion is also represented as providing determiners of errors for the retest method. Finally, let e^2_r be more distinctly the source of error when the retest method is applied, but as being a source of true variance for the other methods. The actual situation is probably not so simple as this, but it is hoped that this much simplicity will contribute to clear conceptions.

Now for some illustrations of actual determiners of the different kinds of variance. These determiners, it must be remembered, are thought of as producing individual differences between scores, either within a single application of a test or between applications or between forms. Among the determiners of individual differences that are consistent from time to time and from one form of a test to another is individual status in some enduring ability, skill, or other trait or traits. These are the things that we wish to measure. Incidental determiners that also belong under por-

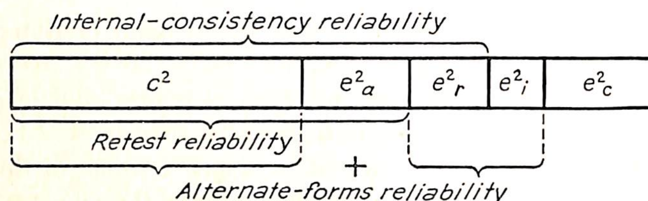


FIG. 17.4.—Proportions of the total-score variance that can be regarded as true variance or as error variance depending upon which type of reliability estimate is made.

tion c^2 in the diagram (Fig. 17.4) are general skill in taking tests, skill in taking this particular kind of test, including the form of item used, and possibly the ability to understand test instructions. These additional sources of variance are only potential. For any given test, the task may require so little understanding or the type of item may be so well known to all examinees that they are practically on a par with respect to these determiners and they consequently would not contribute to individual differences in scores. If they do operate to affect variances, however, they would produce effects in the same directions in odd and even scores, and in so far as individuals do not change in these respects from one administration to another, they would contribute to true variance in all three types of reliability estimate.

Determiners that contribute to error variance in the retest method include temporary conditions, either of the examinee or of the testing environment, including the examiner. The examinee's state of health, fatigue, boredom, emotional condition, and the like may well change from one day to another. Environmental conditions can vary considerably without affecting scores materially, but in so far as they do, such factors

as temperature, humidity, lighting, audibility of instructions or signals, ventilation, and the like, may differ enough to contribute to error variance. There are probably more important changes in the examinee himself. Having taken a certain test, he is not the same individual when faced with the second attempt. The skills and knowledge acquired during the first administration and in the interval between will have their effects upon the second performance. Memory for answers given on the first occasion may lead to repetitions of the same answers the second time and thus contribute to apparent true variance. Awareness of mistakes made in the first attempt, however, leads to changes in responses and hence to error variance. Besides possible improvement during the taking of the test the first time there is possible improvement resulting from transfer effects occurring during the interval between administrations. There are also possible maturational factors, particularly in young children. If learning and maturational effects were uniform for all individuals, or in proportion to their initial positions in the distribution, these determiners would not contribute to error variance. But to the extent that learning and maturational effects differ from person to person, they do add much to error variance. The longer the time interval between test administration, the greater the error contributions. In some tests, continuous loss in reliability occurs as a function of time interval between test and retest. In some psychomotor tests, self correlations of .90 to .96 may be found by the odd-even method, but test-retest correlations with a year interval between may give correlations of approximately .70. Results of this kind were found in testing aviation cadets in the AAF before training and again after aircrew training and perhaps some combat.

Error variance in the alternate-forms method is contributed chiefly by the change in content of the test. Knowledge and skill for dealing with one particular set of items may vary somewhat from the knowledge and skill for dealing with another set of items, and these variations differ from person to person. In addition, depending upon the time interval between administrations of the two forms, some of the determiners of error variance just mentioned for the retest method may also apply to the alternate-forms method. An experiment in the AAF[1] in which the two forms were given in immediate succession and also with four hours of other testing intervening showed no appreciable change in the size of the self correlation. Longer periods might well be expected to have some effect.

If the odd-even technique is used in the split-half method, the changes

[1] Guilford, J. P. (Ed.) Printed classification tests, in AAF aviation psychology research program reports, Report No. 5. Washington, D.C.: Government Printing Office, 1947, 25*f*.

in conditions that may occur during a single administration of a test are rather uniformly distributed over all items in both halves so that their effects would not show up as error variance. There are other ways of splitting tests into halves, however, which may allow more error variance to creep in. If the test is divided by blocks of items, as in odd and even half-pages, or odd and even two-minute trials, or first half against second half, there is room for systematic shifting of conditions. The effects of learning, of temporary changes in mental set (as for speed versus accuracy or as to mode of attack on the items), or of fatigue or motivation, then might contribute to error variance. These are represented in section e^2_i in Fig. 17.4.

The determiners of error that would affect all methods of reliability estimate alike, represented by e^2_c, are such phenomena as fluctuations of attention or memory or of motivation that occur from moment to moment or from item to item. On some tests, guessing is an important contributor to error variance. If a test is so difficult that everyone does considerable guessing (in the extreme case assume that every examinee guessed on every item) the total scores for all examinees approach chance distributions whose variances are very largely error variance. If guessing is a feature in any test, the more difficult the test, the lower its reliability is likely to be. On the other hand, if the test is too easy, the lower the dispersion of scores and the lower the reliability. The smaller the number of alternative responses, the greater the importance of the guessing feature. True-false tests of the same material are less reliable than are four-choice tests, and these, in turn, less reliable than tests of the completion form, other things being equal. The moral of this, of course, is to avoid items with too small a number of alternative responses or to compensate for the greater chance element by making the test longer.

When Different Methods of Estimating r_{tt} Are Preferred.—Preference for one of the three types of reliability estimate depends mostly upon two considerations: type of test and meaning of the statistic, or purpose for which it will be used.

Homogeneous versus Heterogeneous Tests.—Psychological tests can be divided roughly into two classes: homogeneous and heterogeneous. The former are functionally uniform or, strictly speaking, factorially unique. They measure one ability or trait. Very few tests satisfy this definition completely. Some examples are vocabulary, numerical-operations, and perceptual-speed tests. The great majority of tests are factorially complex. Each one measures at the same time a number of different abilities or traits. So far as reliability is concerned, other tests may be considered homogeneous if the items are similar in factorial content. That is, if the

test as a whole measures abilities *P*, *Q*, and *R*, and if each and every item also measures those three abilities, for operational purposes the test may be regarded as functionally homogeneous. An example of this would be an arithmetic-reasoning test or a figure-analogies test. We expect that homogeneous tests shall be internally consistent—we want all parts to measure the same thing; consequently, some form of internal-consistency index is called for.

If a test is heterogeneous, in the sense that different parts measure different traits, we would not expect a very high index of internal consistency. An example of such a test is a biographical-data inventory. This kind of test is composed of questions concerning the examinee's previous life and experiences. Each response to every item is usually validated by correlating it with some practical criterion, *e.g.*, success in pilot training. The reason one response is valid is not necessarily the same as the reason another is valid. They may both predict the criterion and yet correlate zero with each other. The parts of such a test, one randomly chosen half and another, will probably not correlate very high with each other. The test has low internal consistency. An r_{tt} computed in this manner would not do justice to the test. Neither would an alternate-forms r_{tt}, if the forms were developed independently. The only meaningful estimate of reliability for a heterogeneous test is the retest variety. If, by chance, a heterogeneous test were developed, each item of which correlated with a criterion and yet did not correlate with any other item, the internal-consistency reliability would be zero. Yet, the retest reliability might be substantial or high. A biographical-data test of the type referred to above had a characteristic split-half reliability coefficient of about .35 and a retest reliability of about .65. Both of these values are unusually low, but the test had a validity close to .40 for the selection of pilots and consequently was very useful. Such a finding as this, incidentally, is a dramatic demonstration of the fact that the requirement of reliabilities of .90 and above is very unrealistic and that if a selection test proves to be valid we can tolerate its low reliability. Standards for evaluating both validity and reliability coefficients must be relative. No universal limits can be applied.

It is clear from the discussion above that the internal consistency and the stability of the same test need not agree very closely. There can be very low internal consistency and yet substantial or high retest reliability. It is probably not true, however, that there can be high internal consistency and at the same time low retest reliability, except after very long time intervals. If the two indices of reliability disagree for a test, we can place some confidence in the inference that the test is heterogeneous.

High internal-consistency reliability is in itself assurance that we are dealing with a homogeneous test, at least within the broad meaning of the term stated above. Certain of the internal-consistency estimates of r_{tt} such as the Kuder-Richardson (to be described later), which pay attention to item performance, give more assurance of this conclusion than do split-half procedures.

Speed Tests and Power Tests.—Tests are also sometimes roughly categorized as speed tests and power tests. There is no sharp line of demarcation. A genuine power test is one that all examinees have time to finish. It is intended that every examinee shall attempt every item. Achievement examinations are in this category. Speed tests are those in which there is a time limit such that not all examinees can attempt all items. In this category are tests ranging all the way from those in which no one attempts all items to those in which 99 per cent may do so. The latter are so close to the power type that many examiners would be inclined to place them in the power category. If this is the case, it is necessary to specify some percentage of "finishers" necessary to make a power test. There are purely speed tests, in which there is a fixed quantity of work to be done and the score is a time score, or a time score modified by virtue of the number of errors.

It would be out of the question to use the odd-even method of self correlation with a highly speeded test. If no examinee finished and if there were no errors, the correlation of halves would be +1.00 which would have no meaning except that the scorer had counted the numbers of reactions in the two halves correctly. If first and last halves were used, assuming everyone finished the first half and there were almost no errors, all scores for the first half would be about the same and those for the last half would depend upon the rate of work. The correlation would be near zero, for lack of dispersion of the first-half scores.

In fact, any internal-consistency estimate of r_{tt} would be misapplied to a speed test. The errors caricatured above are present to some degree no matter which one of the internal-consistency methods we apply. A retest method will be adequate for many speed tests, except where there is identity of items and hence learning and memory are sources of variance, both true and error, in unknown proportions. For most speed tests, and this includes those in which any appreciable number of examinees fail to reach the last item, an alternate-forms type of reliability estimate is probably best.

A good device to use in the development of new tests is to prepare two equivalent halves and to administer them in immediate succession as two separately timed tests. The correlation between the two halves, inde-

pendently administered, can be treated as we treat the correlation of any other half-scores by the Spearman-Brown prophecy formula in order to estimate the reliability of the full-length test. The comparability of the halves can usually be accomplished by careful construction. Some check upon the adequacy of the efforts is in the comparability of means, standard deviations, and skewness of the two distributions.

Meaning and Use of the Indices of Reliability.—The retest method yields information about the stability of rank orders of individuals over a period of time. A high r_{tt} from this source indicates that persons change very little in status from the first to the second testing, also that the test measures the same functions before and after the interval. A low r_{tt} of this type may mean that individuals have changed in different directions or in the same direction at different rates. Changes of means and of standard deviations will help to interpret the kinds of systematic changes taking place. Plots of scatter diagrams may show whether systematic changes are uniform over the range. These changes we call *function fluctuations of individuals.* If the test measures something different after an interval than before, we have a *function fluctuation of the test.* These changes can be examined by means of correlations of the test with other tests before and after the interval.

There may be some practical reasons for knowing the stability of scores over periods of time and, if so, the retest r_{tt} is the index to use. Usually, the length of time is a factor to be considered. The chief use of this information is in deciding whether to depend upon scores that were obtained in an earlier testing or to administer the same test or a new form to obtain some scores that better describe the individuals right now. As a general policy it would be desirable to establish the principles regarding what kinds of tests yield stable scores, with what kinds of populations, and over what periods of time and what kinds of tests do not.

The meaning of internal consistency was covered in a superficial way in the discussion of homogeneous tests. We will go more thoroughly into the matter shortly in treating the specific methods under this category. This concept probably comes closest to the basic idea of reliability. The methods make an estimate of reliability from a single administration of a single test. The estimate is of an "on-the-spot" reliability. It tells us something of how closely the obtained score comes to the score the person would have made at this particular time if we had had a perfect measuring instrument. For some purposes this information will certainly not be sufficient. It is the kind of reliability that does have meaning in connection with factorial descriptions of tests. These descriptions (see Ch. 18) attempt to depict a test in terms of its component variances, some of which

combine to make up its true variance. It tells us nothing about function stability of persons or of tests.

The alternate-forms estimate of r_{tt} tells us something about function stability in variations of the same test or in different items which are designed to measure the same functions. It indicates how independent the measurements are of the particular items or content used. If the two forms happen to be two halves of the same test, then presumably the kind of items is the same in both (verbal, numerical, pictorial—matching, multiple-choice, completion); only the specific problems change. The alternate-forms r_{tt} may tend to be slightly lower than the internal-consistency r_{tt} but this may mean that it gives a more realistic picture of how accurately the test measures the general traits, ruling out whatever variance is dependent upon the particular content of one form of the test. The two estimates will be almost identical, probably, in power tests of very closely matched content. In power tests, then, the two methods could be used almost interchangeably. In speed tests, as indicated before, the alternate-forms method is the most justifiable approach to reliability estimate.

Internal-consistency Reliability

There are several operations by which an internal-consistency estimate of reliability may be made and there is so much basic test theory bound up with them that we need to give this approach special attention. First, we will consider some more theory.

The Statistical Nature of a Test Composed of Items.—Most tests are composed of items. Most tests are scored by giving credit of +1 for a correct response to each item and a weight of 0 for each wrong answer or omission. The theory about to be explained assumes that kind of a test. Furthermore, it applies best to a power test, in which omissions and wrong answers probably mean inability to master the item. For the time being we will not be concerned with the problem of chance success by guessing. We might assume completion items in which chance factors resulting from guessing are almost nil. The theory will probably apply to situations deviating appreciably from these specifications, enough so that the many conclusions to which it leads will have quite general application.

Item Statistics.—It is convenient to think of each item as a subtest in a larger composite. Each item, then, yields a distribution of scores, with a mean and a standard deviation. According to an earlier discussion of proportions (see Table 9.3), the mean of such a distribution, where the measures are either 0 or 1, is equal to p, the proportion of all who attempt the item who get the right answer. The variance of the distribution is equal to pq, where $q = 1 - p$, and the standard deviation is $\sqrt{pq}$.

The total score on such a test is the sum of part scores. In equation form,

$$X_t = X_a + X_b + X_c + \cdots + X_i + \cdots + X_n \quad \text{(The sum of item scores to make a total test score)} \tag{17.10}$$

where $X_a, X_b, \ldots, X_n$ = scores in items $a, b, \ldots, n$, when there are n items in the test.

The variance of the total test score can be derived from the variances and covariances of the items, according to the principles brought out in the preceding chapter in connection with the variance of sums. Equation (16.19) applied to this particular use would read

$$\sigma^2_t = p_aq_a + p_bq_b + p_cq_c + \cdots + p_iq_i + \cdots + p_nq_n + 2r_{ab}\sqrt{p_aq_ap_bq_b} + 2r_{ac}\sqrt{p_aq_ap_cq_c} + \cdots + 2r_{(n-1)n}\sqrt{p_{(n-1)}q_{(n-1)}p_nq_n} \quad \text{(Total test variance as summation of item variances and covariances)} \tag{17.11}$$

where $p_a, p_b, \ldots, p_n$ = proportion passing items $a, b, \ldots, n$.
$q_a, q_b, \ldots, q_n = 1 - p_a, 1 - p_b, \ldots, 1 - p_n$.
$r_{ab}, r_{ac}, \ldots, r_{(n-1)n}$ = intercorrelations of items.

In abbreviated, summational form, the equation reads

$$\sigma^2_t = \Sigma p_iq_{ji} + 2\Sigma r_{ij}\sqrt{p_iq_ip_jq_j} \quad \text{(Same as formula (17.11) in summation form)} \tag{17.12}$$

where $p_i = p_a, p_b, \ldots, p_n$, in turn.
r_{ij} = correlation between item i and item j, where subscript j is numerically greater than i.

Deductions Derived from the Item-variance Equations.—There are many useful and enlightening inferences that can be deduced from the equation just given. We will consider only the most important ones here.

Relation of Variance to Item Difficulty.—The first thing to be noted is the relation of variance to item difficulty. Remembering that variance means individual differences and the greater the variance, the more we have dispersed individuals in measurement, it can be stated that the item that will produce the greatest dispersion is of median difficulty. It is an item passed by half of the group and failed by half of the group. When $p = q = .5$, the pq product is at a maximum. As p approaches 0 or 1 the variance decreases toward the vanishing point. This has a common-sense explanation. Let us suppose an item that one person out of 100 can answer correctly. This item discriminates one person from each of 99, or makes 99 discriminations. Then, suppose an item that can be passed by two out of 100. This items makes 2 × 98 discriminations, or 196. Continue this to 50, and we get 2,500 discriminations, each one of the 50

who pass it from each one, in turn, of the 50 who fail it. Items of moderate difficulty, then, yield the maximum variance.

Relation of Reliability to Item Intercorrelations.—For the sake of internal consistency, however, large item variances by themselves would mean nothing. If equation (17.12) were limited to the item-variance terms alone, the test would have zero internal consistency; zero reliability of the internal type. This kind of reliability comes entirely from the covariance terms and these are composed of item intercorrelations as well as indices of dispersion. It is only by virtue of their entering into the covariance terms that the item variances contribute to internal consistency. The intercorrelations of the items are the essential sources of this kind of reliability. The larger the item intercorrelations, the greater the internal consistency.

The Effect of Range of Item Difficulty upon Relibiality.—Reliability will be higher when the items are nearly equal in difficulty. A wide range of difficulty is not favorable to reliability (though later we will see that it *is* favorable to validity). The reason is that the appropriate index of item intercorrelation is the ϕ coefficient. Operationally, with items scored as either 0 or +1, their distributions are best conceived as point distributions. If two items differ much in difficulty, the proportions passing the two differ and ϕ consequently is restricted in size. Only when the two items are equal in difficulty can the ϕ between them equal +1 as a maximum (see Ch. 13). Two items very far apart in difficulty might correlate less than .20 even when each measures the same thing and measures it well.

Effect of Item Intercorrelations upon Score Distributions.—There is an interesting bearing of the internal consistency of a test upon the form of distribution of total scores on that test. Imagine a test of 10 items each of exactly median difficulty for the population ($p = q = .5$) and each correlated +1.0 with every other item. A person who passes one item would pass them all and a person who fails one item would fail them all. There would be only two scores possible, 0 and 10. If 20 examinees took this test, the chances are good that their frequency distribution would be like the first diagram in Fig. 17.5. There would be perfect and maximal separation of the two groups. The form of the distribution would be U-shaped. Examples of U-shaped distributions can be found in Hull's book on hypnosis and suggestibility, though they are not so extreme as the one in Fig. 17.5.[1] It is probable that some tests of suggestibility are such that if the examinee responds in the suggestible manner in one trial he will respond similarly in all trials.

[1] Hull, C. L. Hypnosis and suggestibility. New York: Appleton-Century, 1933. P. 68.

If the item intercorrelations are not perfect but high, there will be some moderate scores but there will be a distinct tendency toward bimodality. The second distribution in Fig. 17.5 shows this type of test. With still further reduction in item intercorrelation, the distribution approaches rectangular form, as in the third diagram in Fig. 17.5. With still further reduction in correlation, the distribution approaches normal form, but is

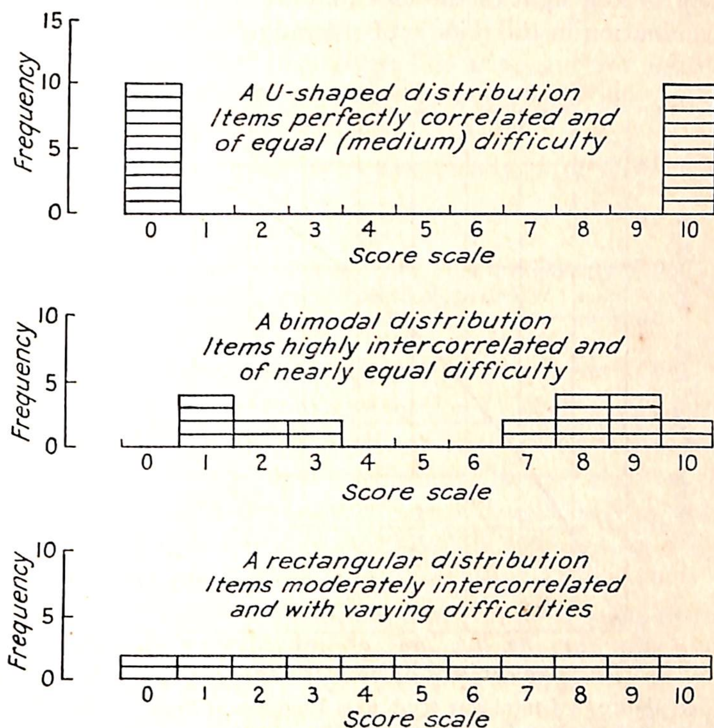


FIG. 17.5. Illustration of the effects of item intercorrelation upon the form of frequency distribution of total test scores.

somewhat platykurtic. A test of zero internal consistency would presumably yield a normal distribution.

These inferences would hold true under the specifications listed: a test of items scored 0 or +1; items of equal difficulty; and items of equal amount of intercorrelation. In practice these conditions are very rarely satisfied in full. The tendencies mentioned, however, may be noted if one is on the lookout for them. It is safe to say that even the most homogeneous tests have average item intercorrelations lower than .30, and in most tests r_{ij} would average nearer .10. Difficulty of items also varies over quite a range; often from a p of .10 to one of .90. Where the range of difficulty is somewhat restricted, however, and item intercorrela-

tions relatively high, platykurtic distributions are not at all uncommon. This discussion should be a restraining influence upon those who insist that tests yield normal distributions of raw scores. If the population distribution is actually normal on the trait measured, then, of course, the platykurtic distribution would distort the facts. If we scale to a normalized distribution, the distortion is corrected. But in this transformation we are likely to lose sight of the fact that we had somewhat better than usual discrimination in the middle of the range.

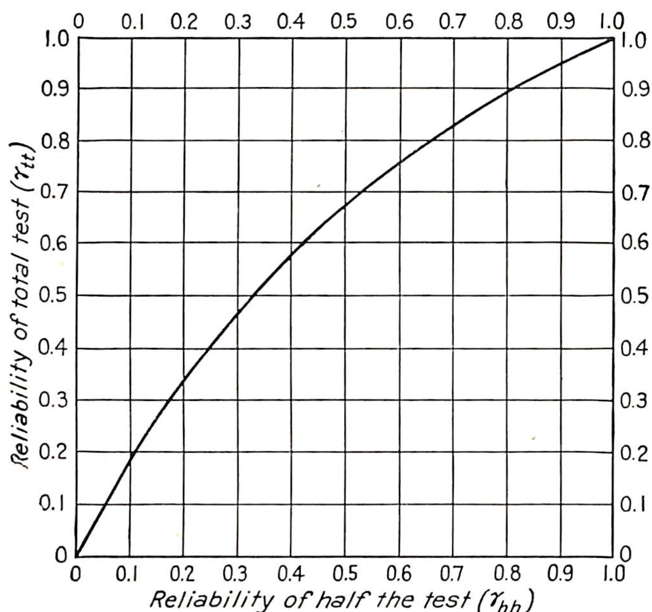


FIG. 17.6.—Reliability of a total-test score as a function of known reliability of a half-test score when the Spearman-Brown formula may be applied.

The Spearman-Brown Prophecy Formula.—The Spearman-Brown formula was designed to estimate the reliability of a test n times as long as the one for which we know a self-correlation. So many times a split-half correlation is known for a test and the correlation of halves is an estimate of r_{tt} for the half test. If it is a homogeneous test, measuring the same trait or traits throughout, the full-length test is not twice as reliable as the half-test, but it is greater and can be estimated by the special Spearman-Brown formula with $n = 2$. If we let r_{hh} stand for the self-correlation of a half-test,

$$r_{tt} = \frac{2r_{hh}}{1 + r_{hh}} \quad \text{(Reliability of a total test estimated from reliability of one of its halves)} \qquad (17.13)$$

When this estimation formula is used, comparability of the halves must be assumed. Comparability is indicated to some degree by the fact of similar means, standard deviations, and skewness of distributions. If comparability is lacking, the reliability of the total test will be underestimated. Since comparability is probably never perfect, an estimate by the use of the Spearman-Brown formula is probably conservative.

Because the split-half method and also the alternate-forms method in the form of two separately timed halves of the same test are so common in practice, the chart in Fig. 17.6 is supplied as an aid in the use of formula (17.13). Since the estimates are rough, in any case, the graphic solution will probably serve for most purposes.

For the general case, in which n could be any ratio of test length to that for which r_{tt} is known,

$$r_{nn} = \frac{nr_{11}}{1 + (n - 1)\, r_{11}} \quad \text{(Spearman-Brown formula for reliability of a test of length } n\text{)} \qquad (17.14)$$

where r_{11} = reliability of the test of unit length.

As a matter of fact, the ratio n in equation (17.14) could be fractional as well as integral. If we knew the self-correlation for a test of 50 items, and we wanted to know the probable reliability for a similar test of 75 items, n would equal 1.5. If we knew the reliability of a test of 100 items and wanted to know approximately the reliability for one of the same kind just half as long, n would be 0.5.

As a matter of interesting information, the Spearman-Brown formula is derived from equations for the correlation of sums. Equations somewhat like (16.24) in the previous chapter have been developed for correlating one composite with another composite, when correlations between parts in each composite and between parts in one composite and parts in the other composite are known. The equation simplifies if the parts have equal variances and equal intercorrelations. The Spearman-Brown formula is such a simplified equation. That is why we have to make the stated assumptions when applying it.

Reliability Estimated from Item-test Correlations.—If we knew the size of the item intercorrelations and if they were uniform in size, or nearly uniform, we could apply the Spearman-Brown formula, letting n equal the number of items, to find r_{tt}.

We would probably not want to take the trouble to determine the intercorrelations among items, but they can be estimated in a manner that is feasible. It has been shown that when item intercorrelations are of about the same magnitude and when items are of approximately equal difficulty,

the average item intercorrelation is equal to the square of the average correlation of items with total score.[1] In a formula,

$$\bar{r}_{ij} = \bar{r}^2_{it} \quad \text{(Relation of average item intercorrelation to average item-test correlation)} \qquad (17.15)$$

where the bars over the r's indicate that they are averages.
r_{ij} = a correlation between item I and item J; a ϕ coefficient.
r_{it} = a correlation between item I and total test score; a point-biserial r.
The item-test correlations are frequently known as a by-product of item analysis. Their mean can be used in the Spearman-Brown formula, which would then read

$$r_{tt} = \frac{n\bar{r}^2_{it}}{1 + (n - 1)\bar{r}^2_{it}} \quad \text{(Estimate of } r_{tt} \text{ from average item-test correlations)} \qquad (17.16)$$

where $\bar{r}_{it}$ = the mean of correlations of items with total test score.

It may be of interest in some connections to estimate the average item intercorrelation from the total test reliability. Using the Spearman-Brown formula with the ratio $1/n$,

$$\bar{r}_{ij} = \frac{r_{tt}}{n + (1 - n)r_{tt}} \quad \text{(Estimate of mean item intercorrelation from } r_{tt}\text{)} \qquad (17.17)$$

where n = the number of items in the test. If n is not large, formula (17.16) tends to overestimate r_{tt}, since r_{it} is a part-whole correlation.

The Kuder-Richardson Estimates of Reliability.—Like the methods just described, the Kuder-Richardson formulas for estimating r_{tt} depend upon item statistics. They were developed because of dissatisfaction with split-half methods. A test can be split into halves in a great many ways, and each split might yield a somewhat different estimate of r_{tt}. The use of item statistics gets away from such biases as may arise from arbitrary splitting into halves.

The Kuder-Richardson methods make the same assumptions as for the use of the Spearman-Brown formula, for the principle is the same as that above where we applied this formula to estimates of item intercorrelation. To repeat, those assumptions call for items of equal, or nearly equal, difficulty and intercorrelation. If the intercorrelations are equal, the items measure the same trait, or traits; there is a functional homogeneity. The most accurate of the practical Kuder-Richardson formulas is[2]

[1] Richardson, M. W. Notes on the rationale of item analysis. *Psychom.*, 1936, **1**, 69–76.

[2] Richardson, M. W., and Kuder, G. F. The calculation of test reliability coefficients based upon the method of rational equivalence, *J. educ. Psychol.*, 1939, **30**, 681–687.

$$r_{tt} = \left(\frac{n}{n-1}\right)\left(\frac{\sigma^2_t - \Sigma pq}{\sigma^2_t}\right) \quad \text{(General Kuder-Richardson formula for estimating reliability)} \qquad (17.18)$$

where n = number of items in the test.
p = proportion passing an item (or responding in some specified manner).
$q = 1 - p$.

It will be recognized, in comparing this formula with equation (17.12), that the numerator term $(\sigma^2_t - \Sigma pq)$ is the sum of the covariance terms in the summation of item variances and covariances used to express the total test variance. The expression Σpq is the sum of the variances of all items. Deducting this quantity from the total test variance, we have left the sum of the covariances. It is in these covariances that the source of *true* variance lies. The ratio of this quantity $(\sigma^2_t - \Sigma pq)$ to the total test variance thus satisfies the basic definition of reliability given in the very first part of this chapter. The factor $n/(n-1)$ is a minor correction that takes into account the item reliabilities.

The steps necessary for the solution of r_{tt} by formula (17.18) are as follows:

Step 1. Determine the variance of the scores for the group of examinees; in other words, find σ^2_t.
Step 2. Determine for every item the proportion passing it (p) and the proportion failing it (q).
Step 3. Determine the variance of each item (the product pq). Sum the pq products for all items.
Step 4. Substitute the known values in formula (17.18) and solve.

A Shorter Approximation to the Kuder-Richardson Reliability.—If we are justified in making the assumption that all items in the test have approximately the same degree of difficulty, we may use a formula that is much less demanding of information.[1] It reads

$$r_{tt} = \left(\frac{n}{n-1}\right)\left(\frac{\sigma^2_t - n\bar{p}\bar{q}}{\sigma^2_t}\right) \quad \text{(An approximation formula for the Kuder-Richardson reliability)} \qquad (17.19)$$

where $\bar{p}$ and $\bar{q}$ = average proportions of passing and failing examinees for each item, respectively.

[1] In the derivation of their formulas, Kuder and Richardson also find that even formula (17.18) depends upon the assumption of approximate equality of difficulty of items. But Brogden has shown empirically that variation in difficulty of items over very wide ranges does not lead to appreciable bias in the estimation of r_{tt} by formula (17.18). Brogden, H. E. The effect of bias due to difficulty factors . . . on the accuracy of estimation of reliability. *Educ. & Psychol. Meas.*, 1946, **6**, 517–520.

The values of $\bar{p}$ and $\bar{q}$ can be obtained without counting successes and failures for every item, for the average p is equal to the mean of the total scores divided by n, and the average q is $1 - \bar{p}$. From these facts, the formula can be simplified to

$$r_{tt} = \frac{n\sigma^2_t - \overline{RW}}{(n-1)\sigma^2_t} \quad \text{(Alternate to formula (17.19))} \quad (17.20)$$

where $\bar{R}$ = average number of right responses.
$\overline{W}$ = average number of wrong responses (or $n - \bar{R}$).
$\bar{R}$ is of course the mean of the total scores.
In more familiar symbols,

$$r_{tt} = \frac{n\sigma^2_t - M(n - M)}{(n-1)\sigma^2_t} \quad \text{(Another substitute for formula (17.19))} \quad (17.21)$$

It should be said that all the Kuder-Richardson formulas, indeed all the internal-consistency formulas which depend upon a single administration of a test, probably underestimate the reliability of a test. Of all these formulas, (17.18) should usually come closest to the correct value of r_{tt} under the conditions of testing prevailing. Although some of these formulas get away from appearance of item statistics in them, it should not be forgotten what assumptions are implied. They do not apply to speed tests, including, in fact, most time-limit tests.

Several other variations of the formulas have been proposed to meet special requirements. Hoyt suggests a formula convenient for use with raw data, a formula not requiring the computation of a mean or a variance.[1] It reads

$$r_{tt} = \frac{n}{n-1}\left[\frac{N\Sigma X^2 + \Sigma I^2 - \Sigma X(\Sigma X + N)}{N\Sigma X^2 - (\Sigma X)^2}\right]$$

(Hoyt's version of the Kuder-Richardson formula (17.18)) (17.22)

where X = a score in the test.
I = number of right answers for an item.
N = number of scores.
n = number of items in test.

For the test in which items are weighted differently, Dressel has provided the following variation:[2]

[1] Hoyt, C. J. Note on a simplified method of computing test reliability. *Educ. & Psychol. Meas.*, 1941, **1**, 93–95.

[2] See Dressel, P. L. Some remarks on the Kuder-Richardson reliability coefficient. *Psychom.*, 1940, **5**, 305–310.

$$r_{tt} = \left(\frac{n}{n-1}\right)\left(\frac{\sigma^2_t - \Sigma w^2_i p_i q_i}{\sigma^2_t}\right) \quad \text{(Kuder-Richardson formula when items are weighted)} \quad (17.23)$$

in which w_i = weight for item I. Dressel also provides other formulas to apply when scoring formulas are used, weighting wrong responses and omissions differently.

The Rulon Method of Estimating r_{tt}.—It was mentioned earlier that Rulon had developed a method of computing the standard error of an obtained score, $\sigma_{t\infty}$, from differences in scores on two halves of a test. Because of the relations between r_{tt} and $\sigma_{t\infty}$, the same approach leads to another kind of estimate of reliability. It is usually applied to halves of the test in a single administration and hence comes under the category of an internal-consistency reliability, but it could be applied to alternate forms. It might even be applied to a retest approach, under special conditions.

Because $\sigma^2_{t\infty}$ measures the amount of error variance, an estimate of r_{tt} is given by the formula

$$r_{tt} = 1 - \frac{\sigma^2_{t\infty}}{\sigma^2_t} \quad \text{(Reliability by the Rulon formula)} \quad (17.24)$$

where $\sigma^2_{t\infty} = \Sigma d^2/N$, as in formula (17.9).

Rulon's formula[1] is especially applicable when an IBM test-scoring machine is available, for this instrument can be so adjusted as to yield a difference between odds and evens for each examinee. The method is subject to the same restrictions as for any split-half procedure. It should be noted that *the formula gives the reliability of the total test scores and not of the halves;* so the Spearman-Brown formula need not be applied. If the Rulon difference formula should be applied to differences between scores on two forms, the reliability coefficient thus estimated applies to a test of twice the length of either form. A correction to the reliability wanted for each form can be made by substituting .5 for n in formula (17.14).

A Summary of Internal-consistency Reliability.—Internal-consistency reliability is most appropriately applied to homogeneous tests; *i.e.*, tests composed of equivalent units—equivalent in several respects. The parts (usually items) all measure the same trait, or traits, to about the same degree. The total variance of a test can be conceived as a sum of the variances and covariances of its parts. The true variance of a test is contributed by its covariances to which both the item variance and item

[1] Rulon, P. J. A simplified procedure for determining the reliability of a test by split-halves. *Harv. educ. Rev.*, 1939, **9**, 99–103.

intercorrelations are important contributors. Internal-consistency reliability is greatest when:

1. The item intercorrelations are greatest.
2. The variance of items is greatest. This is when the proportion passing an item is .50.
3. The items are of equal difficulty. Then the item intercorrelations are at a maximum.

In estimating an internal-consistency r_{tt}, most methods rest upon the assumptions of equivalence of parts in the sense of equality of difficulty and equality of intercorrelation. If these conditions are not satisfied, estimates of r_{tt} may still be made, but the farther the departure of the situation from these specifications, the more is r_{tt} likely to be underestimated.

Achieving Internal Consistency in a Test.—Since the key to internal consistency is item intercorrelation, the way to improve this type of reliability is to increase this feature of the test. The best practical index of item intercorrelation is the correlation of an item with total score on the test. In formula (17.15), we saw that under the right conditions the average item intercorrelation in a test equals the square of the average correlation of items with total score. Under somewhat liberal conditions, item-test correlations indicate the homogeneity of the item with the rest of the items and its right to membership in the test. The effectiveness of item-test correlation as an index of internal consistency depends upon how well we have defined the trait to be measured and how well we have devised items to measure it. It is presumed that in striving for a homogeneous test we want it to measure as nearly as possible one unique trait and one only.

Some Preliminary Precautions in Item-test Correlation.—A number of cautions and restrictions should be mentioned in connection with this process. In the first place, we should be informed as to what the total provisional scores represent. It has too often been assumed that because in the end we have only items that do possess considerable internal consistency or correlation with the same criterion, they are diagnostic of some unitary trait. This is not necessarily correct. The total collection of items, although centering about some hypothetical unitary trait like introversion-extroversion, is usually a measure simultaneously of *several* real variables in personality. Unless a factor analysis or something equivalent has been made to establish the unity of the trait, the variable measured is probably complex. Even if it is shown that the items correlate with the total, there may be subclusters that correlate well among themselves but not so much with items in other clusters. One item might correlate with the total score because it has in common with

the total score validity for trait P, but another so correlates because of its relation to trait Q. Nor does the fact that the items do not correlate with test scores outside this one or the fact that total scores do not correlate with one another necessarily point to the fact that each test measures a single variable. The route to the measurement of unitary variables cannot be taken exclusively by way of tests of internal consistency.[1]

Another minor difficulty is that the item itself helps to determine the total score, and we are dealing with correlation of part with whole. When the test is a long one, more than 50 items, this is of trivial consequence. It calls for some kind of correction or allowance when the number of items is 20 or less.

Methods of Determining Internal Consistency of Items.—There are many ways of determining how well a single item agrees with the total score in the kind of discriminations it makes among individuals. Most of these procedures are correlation methods. The correlation methods that have been used in this connection include the biserial r, point-biserial r, ϕ coefficient, and tetrachoric r. Whatever index of correlation is used, one variable is of necessity a dichotomy, as a rule. It is a division of the sample into those who pass the item and those who fail it. In the case of interest and temperament items, the division is in terms of those who give one of the responses versus those who give other responses. The total test score is a continuous variable, but for convenience it is usually artificially dichotomized for the sake of simplicity of numerical operations.

Previous comments in several places have indicated that the pass-fail variable for a single item should be regarded as a genuine dichotomy, a point distribution. This would rule out the use of the biserial r and the tetrachoric r as indices of item-test correlation. If it is one's desire to know the degree of correlation between whatever the item measures and what the total score measures, then these two coefficients would be quite in order. But since we are dealing with a prediction problem in the use of test items, and since we can only predict from one of the two categories, pass or fail, the realistic index of the functioning of the item is the point biserial. If the ϕ coefficient is substituted for the point biserial and if we wish to know the point-biserial r, a correction for coarse grouping could be made in one variable. Ordinarily, we are concerned only about the rank order of correlations of items with total score and we can select items of highest correlation without making any correction. The ϕ coefficients would be in approximately the same rank order as the point biserials. Furthermore, it is quite customary to reduce the total-score variable to

[1] Sletto, R. F. Construction of personality scales by the criterion of internal consistency. Minneapolis: Sociological Press, 1937.

virtually a point distribution by leaving out the middle of the sample, using only the highest and lowest quarters. With this step taken, the ϕ coefficient is the most realistic index of correlation to use. It can be used when the two groups on total score are highest and lowest halves, but the size of the coefficients will then be somewhat smaller.

In dichotomizing the total-score variable, whatever choice is made as to percentages in the two groups, the convenience of having multiples of 100 in each of the two groups probably outweighs other considerations. There is some statistical evidence that the index of correlation will be most sensitive when the highest and lowest 27 percentages are selected. To secure the desired numbers of 100 in each category, we would need to examine about 370 for an item analysis. There is probably little loss in efficiency of procedure by deviating from 27 per cent, even to the extent of using upper and lower thirds. If a full 400 have been examined, it might be better to include *all* the sample, 200 in each group, but upper and lower quarters would be preferred to 27 percentages for the convenience of using round numbers.

The Phi Coefficient as the Index of Item Consistency.—Experience has shown that the ϕ coefficient is among the most satisfactory indices of item consistency from several points of view. The computational aids, as described below, reduce the effort to a minimum which some other methods equal but do not go below. Its direct relation to chi square provides automatically a test of minimal significant coefficients for any size of sample.

A well-controlled study was made of a number of the methods, including biserial and point-biserial r's, an ordinary tetrachoric r, the Flanagan tetrachoric r (based upon extreme 27 percentages),[1] and two applications of ϕ (with higher and lower halves and with extreme 27 percentages). All methods were applied to the results from two administrations of the same test composed of 68 items to two samples of 400 aviation students each.[2] The different methods were compared in several ways. Questions asked included: "How consistent is the same coefficient for the same item from one sample to the other?" "How much dispersion among items does each index provide?" The index giving the greatest dispersion may be regarded as perhaps most sensitive to differences in correlation. "Which one gives coefficients that agree best with those from other methods?"

Table 17.2 presents the answers to these questions in brief form. The

[1] Flanagan, J. C. General considerations in the selection of test items, etc. *J. educ. Psychol.*, 1939, **30**, 674–680.

[2] Guilford, J. P. Printed classification tests. P. 30.

phi coefficient from extreme 27 percentages leads the list for dispersion of values and for consistency from sample to sample. It is interesting to

TABLE 17.2.—A COMPARISON OF THE EFFECTIVENESS OF SIX INDICES OF ITEM-TEST CORRELATION

Index	Statistics				Rank orders		
	Mean	σ	r_{tt}	Mean correlation with other indices	σ	r_{tt}	Mean correlation
Phi, 27%	.42	.21	.91	.914	1	1	5
Phi, 50%	.29	.14	.87	.922	6	4	3
Flanagan r	.46	.20	.87	.916	2	4	4
Point biserial r	.35	.15	.88	.934	5	2	1
Biserial r	.47	.18	.87	.910	4	4	6
Tetrachoric r	.46	.19	.79	.926	3	6	2

note that the point-biserial r gave results which correlated highest on the average with results from all other indices. This may or may not support the idea that this index is a kind of common denominator because it is the most realistic one. The differences in average correlations of the indices are very small. The entire table of intercorrelations tends to show greatest affinity for methods that use the same data; for example, a correlation of .98 between the 27 per cent phi and the Flanagan r, and a correlation of .96 between the 50 per cent phi and the tetrachoric r.

The mean sizes of coefficients are in line with expectations. The smallest mean would be the phi with 50 per cent division and the next smallest would be the point-biserial r. The Flanagan r should have the same mean as the tetrachoric r, and both nearly the same as the biserial r. It is noteworthy that the phi with 27 per cent extremes gives an average coefficient that is not very far from the three just mentioned. This finding loses some of its significance, however, when we consider some of the peculiarities of phi.

Some Peculiarities of Phi as an Index of Item Correlation.—In Ch. 13 it was pointed out that phi has restrictions which render it different from other coefficients of correlation. To some extent the point-biserial r shares some of these peculiarities. The size of phi is limited by the evenness or unevenness of means in the two variables correlated, and this is indicated by the two marginal proportions p and p' (or q') in the fourfold table. In the item-test correlation situation, the one variable is evenly split, with $p' = .50$. Only items at median difficulty can correlate per-

fectly with total score. The further item difficulty departs from .50, the smaller the correlation will have to be.

But these very peculiarities of ϕ make it a realistic index of consistency. When we want to increase reliability of a test, we strive toward the goals mentioned above—median difficulty level and equality of difficulty among items. Thus, selection of items on the basis of phi favors optimal difficulty conditions as well as agreement in the sense of correlation. The other coefficients, except the point biserial, are not, seriously at least, affected by item difficulty. They tell of agreement of what is measured by the item with what total score measures. When the aim is not merely for reliability, a knowledge of correlation freed from the influence of difficulty level may be very important. In these cases some other index of correlation is superior to ϕ. If ϕ is used in such situations, allowances should be made for the limitations imposed upon it by difficulty.

Allowance for Spurious Part-whole Correlation.—An item correlated with a total score of which the item is a part gives a somewhat spurious correlation. When the test is not very long this feature deserves some attention, particularly where there is a question of retaining items with low coefficients. A rough idea of the extent of this spuriousness can be gained by knowing that if the items are of equal difficulty, any item not functionally related to the total score would correlate to the extent of $1/\sqrt{n}$, where n is the number of items. When n is 50, a completely spurious item-test correlation would be .14; when n is 25, r could be .20; and when n is 10, r could be .32. Since we usually try not to select items with correlations less than about .30, entirely spurious correlations probably do not lead to erroneous selection of items, but the element of spuriousness may well boost an item correlation above significance limits or even above .30 on occasion. One device to avoid spuriousness in item correlations is to divide the test into halves and use each half as the criterion or total score with which to correlate items in the other half. If there is high correlation between halves, this procedure is a good solution.

A Special Formula for Phi in Item-test Correlation.—When one of the variables in the computation of a ϕ is evenly divided, and in item-test correlations the total-score variable is usually so divided, the formula for computing ϕ reduces to the very convenient form

$$\phi = \frac{p_u - p_l}{2\sqrt{pq}} \quad \text{(Phi computed when one variable is evenly divided)} \qquad (17.25)$$

where p_u = proportion of upper total-score group that responds correctly (or in some specified manner) to the item.

p_l = proportion of lower total-score group responding in same manner.

p = proportion of two subgroups combined that react in same manner and is given by the relation

$$p = \frac{p_u + p_l}{2}$$

$q = 1 - p.$

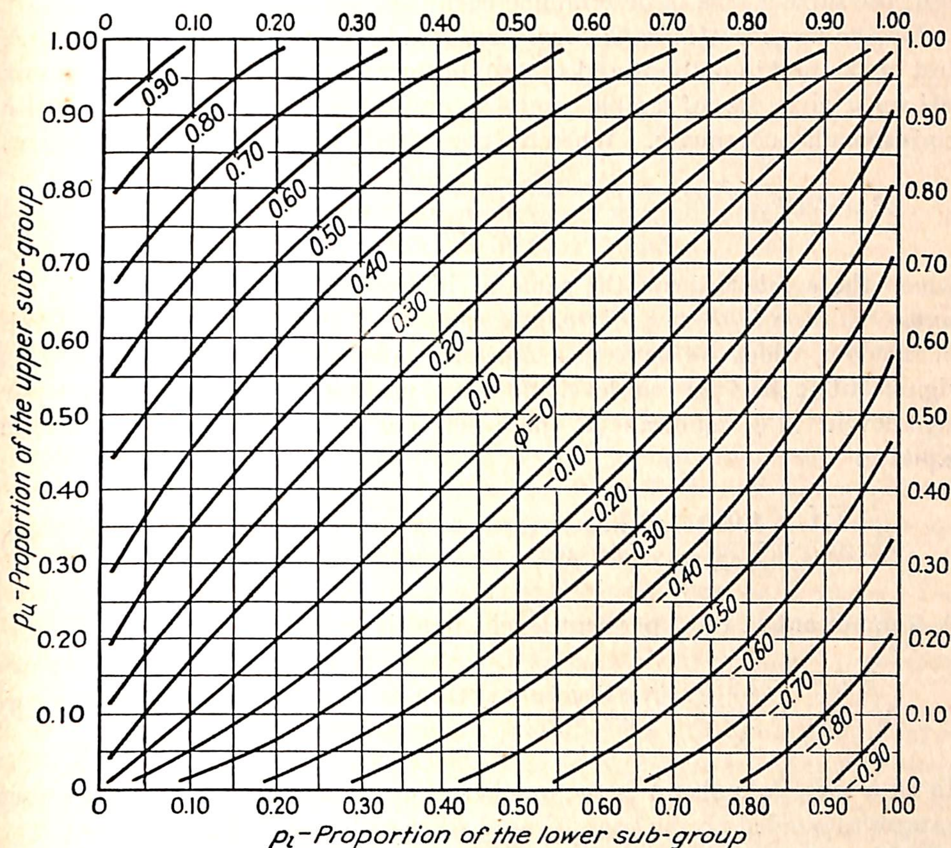


FIG. 17.7.—An abac for graphic estimates of the phi coefficient when one variable has an even division of cases in its two categories. If the proportion of the upper criterion group passing an item is 0.65 and the proportion of the lower group passing it is 0.30, ϕ is found at the intersection of the horizontal line at level 0.65 and vertical line at position 0.30. It is midway between the lines for $\phi = 0.30$ and $\phi = 0.40$; therefore, the ϕ we are looking for is 0.35.

An abac developed by the author for the solution of ϕ when this formula applies is given in Fig. 17.7.[1] In using the abac, the only information we

[1] Guilford, J. P. The phi coefficient and chi square as indices of item validity. *Psychom.*, 1941, **6**, 11–19.

need from the data are p_u and p_l. Looking up the ordinate and abscissa values corresponding to p_u and p_l respectively, for a given item, we find that the intersection of the lines, horizontal and vertical from those points, locates ϕ, which can be estimated to the second decimal place. We interpolate, when necessary, between the curved lines, each of which represents a constant value of ϕ, by considering the perpendicular distance at that place. Thus, once the count of responses has been made for items in a test, the further task of determining each item's ϕ is a rapid process.

Determining the Minimum Significant Phi for Item-test Correlations.—A test of the null hypothesis and the establishment of the lowest significant phi for a given size of sample can be accomplished through the use of the corresponding chi square. When the two criterion groups are of equal size,

$$\chi^2 = N\phi^2 = \frac{N(p_u - p_l)^2}{4pq} \quad \text{(Chi square in a fourfold table in which one variable is evenly divided)} \quad (17.26)$$

where the symbols mean the same as in formula (17.25) and N *is the number of cases in the two subgroups combined.* For the case of one degree of freedom, which we have in a fourfold table, a chi square of 3.841 is significant at the 5 per cent level and one of 6.635 is significant at the 1 per cent level. A ϕ significant at the 5 per cent level, therefore, would be equal to

$$\sqrt{\frac{3.841}{N}} = \frac{1.960}{\sqrt{N}} \quad \text{(A } \phi \text{ significant at the 5 per cent level in a fourfold table)} \quad (17.27)$$

A ϕ significant at the 1 per cent level would be equal to

$$\sqrt{\frac{6.635}{N}} = \frac{2.576}{\sqrt{N}} \quad \text{(A } \phi \text{ significant at the 1 per cent level in a fourfold table)} \quad (17.28)$$

In both these formulas N is the number of cases in both upper and lower categories divided on the basis of total score.

Table 17.3 provides a number of the significant and very significant minimum phi coefficients for samples varying in size from 40 to 2,000. If the practice of restricting categories to multiples of 10 and 100 or of 25 or 50 is followed, these tabled values will indicate the required minimal phi coefficients. Ordinarily, in studies of internal consistency, the selection of items will be from among those that correlate with total scores well above the significance levels. When items are correlated with an outside criterion, however (see Ch. 18), resulting ϕ's near the confidence limits are the rule rather than the exception, hence Table 17.3 would be more likely to come into use in that connection.

TABLE 17.3.—PHI COEFFICIENTS SIGNIFICANT AT THE FIVE AND ONE PER CENT LEVELS FOR DIFFERENT SIZES OF SAMPLES

N	5 per cent level	1 per cent level	N	5 per cent level	1 per cent level
40	.310	.407	250	.124	.163
50	.277	.364	300	.113	.149
60	.253	.333	350	.105	.138
70	.234	.308	400	.098	.129
80	.219	.288	500	.088	.115
90	.207	.272	600	.080	.105
100	.196	.258	700	.074	.097
125	.175	.230	800	.069	.091
150	.160	.210	1000	.062	.081
200	.139	.182	2000	.044	.058

Other Indicators of Item Consistency.—From the mention of chi square in formula (17.26) above, it follows that this statistic is another possible indicator of item consistency. It carries its own test of significance. Its size is dependent upon N, however, and only when working with samples of constant size would items be ranked in proper order for correlation with total score. Since ϕ is independent of N, it is a better general-purpose index. It is also easier to compute. Chi square has an advantage of being useful when there are more than two categories of response. It would give a single index of consistency for the item as a whole. But we could apply ϕ to each response in turn and find out where the greatest significance lies and what responses could possibly be combined for scoring purposes.

Another sampling statistic related to chi square when there is one degree of freedom is the t ratio. The alert reader will have recognized the constants in formulas (17.27) and (17.28) as the familiar t ratios that are significant at the 5 and 1 per cent levels for large samples, 1.960 and 2.576. Having as data, p_u and p_l, as defined for formula (17.25) above, we could apply the t test for significance of their differences. Not only the relative sizes of t for different items but also whether or not they reached the standard levels of significance could be determined. Like chi square, however, the size of t depends upon the size of sample and only when N is constant will the t ratios place items in correct rank order for correlation with total score. Mosier and McQuitty have prepared an abac from which the t ratio can be read when we know the two required proportions.[1]

[1] Mosier, C. I., and McQuitty, J. V. Methods of item validation and abacs for item-test correlation, etc. *Psychom.*, 1940, **5**, 57–65.

The Problems of Incomplete Data.—When an experimental test is administered with a time limit such that appreciable numbers of examinees do not attempt all items, the analysis of the items at the end of the test becomes a problem. This is true both for the determination of difficulty of items and for item-test correlations. It is probable that only the more able examinees reach the last items. Whereas the analysis of early items in the test is based upon the total sample, the analysis of the last items is then based upon a more restricted sample of higher average ability.

First, there is the decision as to whether the analysis of the terminal items shall be based upon the total sample or upon only those who attempted the items. If the former is the chosen alternative, many of the "failures" for the terminal items are merely "did not have a chance to attempt for lack of time" propositions. It is obvious that attempting an item and failing is not the same thing as not having been administered the item at all. If the second alternative is chosen, we have a different population for the terminal items than for the early items, and the population shifts as the end is approached. Difficulty values and correlations are bound to be affected by this changing of population.

There are several possible solutions to this problem. The best one is that experimental tests to which item analysis applies should be given with sufficient time for practically all examinees to finish. There may be administrative difficulties in this, since some examinees take at least twice as long as others, but these difficulties are not insurmountable. One device would be to place at the end of the test some very difficult items that probably will not be used but which would serve as "busy work" for those who reach them. Ingenuity will suggest other devices. There is probably no fully satisfactory way of dealing statistically with items that appreciable numbers do not attempt.

Some Special Problems in Reliability

Like all coefficients of correlation, r_{tt}, however estimated, must be interpreted in a relativistic manner. Its size depends upon many conditions under which it is obtained experimentally. Some of the more important conditions and considerations will be mentioned in what follows.

Reliability in Different Ranges of Measurement.—Like intercorrelations of different variables, self-correlations are affected by the range of ability or of a trait present in the population sampled. The narrower the range, the smaller r_{tt} tends to be. This can be seen mathematically if one examines formula (17.4), where r_{tt} is given as equal to $1 - \sigma^2_{t\infty}/\sigma^2_t$. If the standard error of the obtained score remains constant regardless of the range of ability in the sample, we see that if the range, as measured

by σ_t, decreases, the denominator σ^2_t decreases, the ratio $\sigma^2_{t\infty}/\sigma^2_t$ increases, and r_{tt} decreases. This is why some test users prefer to know $\sigma_{t\infty}$ rather than r_{tt} concerning a test, since it is probably more stable from population to population. It is another good reason we should not speak of *the* reliability of a test. Figure 17.8 illustrates how in a restricted sample (small square) the same scatter of points gives a relatively wider spread and hence a lower correlation. Restriction is not ordinarily as clear cut as this in practice, but the principle is the same.

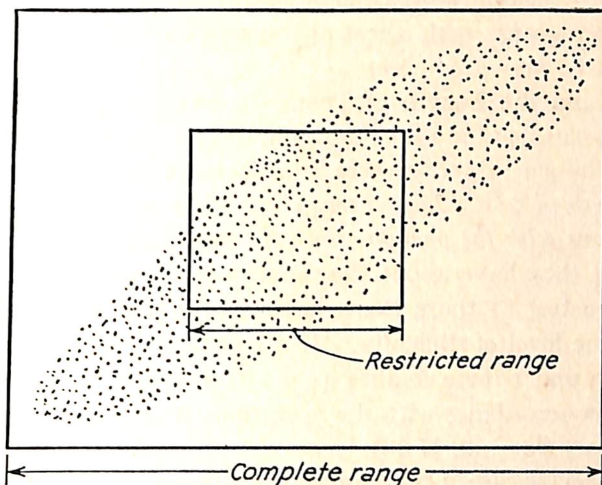


FIG. 17.8.—Illustration showing an extreme instance of curtailment of range. The correlation for the cases within the smaller rectangle will be much smaller than the correlation of all cases within the larger rectangle.

If we wish to estimate the reliability coefficient in one range from the known reliability in another range, the following formula may be used. It assumes equal standard error of obtained scores in both ranges.

$$r_{nn} = 1 - \frac{\sigma^2_o(1 - r_{oo})}{\sigma^2_n} \qquad (17.29)$$

(Estimation of r_{tt} in a population of one dispersion from that in another similar population of different dispersion)

where σ_o = standard deviation of the distribution for which the reliability coefficient is known.

σ_n = standard deviation of the distribution for which the reliability is not known.

r_{oo} and r_{nn} = reliabilities in the two respective distributions.

If we know that a more limited group has a standard deviation of 8.0 and a reliability coefficient of .85 for a test, what will be the reliability coefficient in a more variable group whose σ is 10.0? Applying formula (17.29),

$$r_{nn} = 1 - \frac{8^2(1 - .85)}{10^2}$$
$$= 1 - \frac{(64)(.15)}{100}$$
$$= 1 - .096$$
$$= .904$$

There is a standard-error formula for this estimated r_{nn}, but because standard errors for coefficients of correlation are of little or no value for such high r's as reliability coefficients usually are and since one can then almost always dispense with a test of the null hypothesis, such standard errors will not be presented here.

Reliability and the Length of Test.—It was indicated in connection with the split-half method that the whole test is more reliable than either half and that in general terms there is an increase in reliability going with increased length of test. *This is true if the additional items added to a test are homogeneous with the ones to which they are added.* By *homogeneous* we mean that they have about the same intercorrelation with the items already in the test as those items have among themselves and possess about the same level of difficulty. If a test is lengthened to n times its present length under these conditions, we have a right to expect a change in reliability in accordance with the Spearman-Brown equation which was given previously (formula 17.14).

To take a specific case, a test containing 50 items has a reliability of .80. What would be its reliability if we add 100 more items like the 50 we have? The solution is

$$r_{33} = \frac{3(.80)}{1 + (3 - 1)(.80)}$$
$$= \frac{2.40}{2.60}$$
$$= .92$$

We can also predict reliability in a shorter test from the known reliability in a longer one. If we make a test one-half as long, n becomes .5, and so on. We can let n be any ratio we please. If we increase the test's length 45 per cent, n becomes 1.45.

Lengthening a Test to Attain a Certain Desired Reliability.—We can use the Spearman-Brown formula in reverse. If we know the reliability of a short test is .75, we can ask how long the test would have to be to attain a reliability of .90. If we solve the equation of the Spearman-Brown formula to find n, it becomes

$$n = \frac{r_{nn}(1 - r_{11})}{r_{11}(1 - r_{nn})} \quad \text{(Estimation of length of test required for a given reliability)} \qquad (17.30)$$

Substituting the known values in this equation, we have

$$n = \frac{.90(1 - .75)}{.75(1 - .90)}$$
$$= \frac{.225}{.075}$$
$$= 3.0$$

The test with $r_{11} = .75$ would have to be three times as long to attain a reliability of .90.

Any other level of reliability, larger *or smaller*, in which we are interested can serve as r_{nn}, and the necessary n ratio can then be computed. Experience will show that some tests of low reliability cannot reach some desired high reliability without being made indefinitely long, or so long as to be impractical. Others will exhibit promising improvements in reliability with a moderate amount of extension. The formula is useful in this respect, that it helps decide upon rejection or extension of tests, or it is useful in cases in which a test is already too long for comfort and we need to decide whether shortening it would sacrifice too much in reliability.

Reliability of Ratings and Other Judgments.—Many of the statistics described in connection with test scores also apply fairly well to human judgments of various kinds. The judgments may be in the form of rank order, rating-scale evaluations, paired-comparisons scaling, judgments in equal-appearing intervals, and the like. We can correlate the same observer's judgments obtained at two different times, or we can assume that similar judges are interchangeable and so intercorrelate their evaluations. We can pool judgments for two comparable groups of observers and correlate them so long as they apply to the same objects or persons. Experience has shown that with due cautions these applications may be made with meaningful results. Every coefficient must, as usual, be interpreted in the light of the manner in which it was obtained. Even the Spearman-Brown formula has been shown to apply, as, for example, in the pooling of judgments from two observers, which yields increased reliability in a manner found for the doubling of a test in length. The comparability of judges must be true here just as the comparability of items must be true in applying this formula to the change in length of test.

Exercises

1. The following reliability coefficients were presented for a certain test:

Split half	.96	Retest after 1 month	.91
Alternate form	.94	Retest after 2 years	.86

Are these coefficients reasonable? Discuss.

2. In six tests, the following correlations were found between halves composed of comparable items: .43, .55, .66, .74, .86, and .94.

a. Determine the reliability coefficient for the full-length tests.

b. Determine the index of reliability in each case.

3. In a certain test, the sum of the squared differences between scores on two comparable halves equaled 285. $N = 50$, and $\sigma = 8.5$. Find the coefficient of reliability for the total scores and the standard error of estimate of the obtained scores.

4. In a test of 55 items, the standard deviation of the total scores was 7.5. The sum of the variances of the items was 9.8327. Find the reliability coefficient.

5. Another test of 150 items has a standard deviation of total scores equal to 24.4 and a mean of 94.2. Find the reliability coefficient, assuming that the items are approximately equal in difficulty.

6. In four tests, the reliability coefficients were .65, .76, .87, and .94. Determine $r_{t\infty}$ and $\sigma_{t\infty}$ in each case, assuming that $\sigma_t = 10.0$.

7. A test has $\sigma_t = 7.2$ and $r_{tt} = .86$. In another group from the same population, σ_t is 6.0. What will the reliability become? In still another group, $\sigma_t = 9.0$. What will r_{tt} become?

8. Complete the following table by performing the necessary computations:

r_{11} \ n	1.5	2	4	6	10	20
.30	.39					.90
.70			.90	.93		
.90	.93	.95			.99	

9. For the data in the last exercise, plot on graph paper the increase in r_{nn} (on the ordinate) as n increases (on the abscissa) for each value of r_{tt}. Draw some general conclusions from the table or from the diagram.

10. Complete the following table by computing the necessary n's:

r_{11} \ r_{nn}	.65	.75	.85	.95
.30				
.50				
.70				
.90				

11. Determine for the items in Data 17*A* one or more of the indices of item-test correlation mentioned in this chapter. Tell something about the significance of each index, and draw any other conclusions that suggest themselves.

12. Give evidence of the degree of item-test correlation of responses to the items in Data 17*B*. Are all responses equally diagnostic? Discuss.

DATA 17*A*.—FOR 10 TEST ITEMS ARE GIVEN THE PROPORTIONS OF THE INDIVIDUALS IN UPPER AND LOWER QUARTERS OF A CRITERION GROUP WHO PASSED EACH ITEM AND THE NUMBER OF CASES IN THE TWO SUBGROUPS COMBINED (N)

Item	p_u	p_l	N
1	.84	.64	50
2	.90	.56	50
3	.80	.45	100
4	.75	.58	100
5	.15	.21	200
6	.45	.27	200
7	.62	.52	400
8	.96	.90	400
9	.47	.40	1000
10	.56	.61	1000

DATA 17*B*.—PROPORTIONS OF TWO CRITERION SUBGROUPS WHO RESPONDED IN ONE OF THREE WAYS TO TWO QUESTIONNAIRE ITEMS
($N = 500$)

Question 1. Do you daydream frequently?

Group	Yes	?	No
Low cycloid..........	.46	.09	.45
High cycloid..........	.71	.07	.22

Question 2. Do you consider yourself less emotional than the average person, *i.e.*, less easily upset?

Group	Yes	?	No
Low cycloid..........	.55	.04	.41
High cycloid..........	.30	.04	.66

CHAPTER 18

VALIDITY OF MEASUREMENTS

While most of the comments in this chapter will be about the validity of tests, the problem of validity applies to all kinds of measurements. Most of what is said about validity of tests applies to other methods of evaluation and measurement.

Problems of Validity

It is usually easy enough to apply a metric instrument and to obtain some numerical data. In the physical sciences the meaning of numbers that are used to describe phenomena is usually well established. The values stand for degrees of electrical resistance, pressure of a gas, or mass of a particle. In the social sciences, however, the connection between a number and the thing, or things, for which it stands is not nearly so obvious. Nor is the situation helped very much or the problem solved by conjuring a name for a supposed variable that the numbers stand for. There is said to be a country in which it is regarded as bad taste for anyone to question whether a certain test measures trait *X* if the distinguished psychologist who invented the test says it measures trait *X*. There are other, supposedly more enlightened countries, unfortunately, in which the same attitude exists to some degree in some quarters. The problem would not be so serious if conclusion after conclusion about supposed underlying properties were not built upon the evidence of measurements which may not, after all, have much to do with those properties. There may even be considerable question, also, about the *existence* of the properties.

Types of Validity.—The question of validity, of a test or of any metric instrument, has many facets and it requires clear thinking not to be confused by them. In crudest terms, we say that a test is valid when it measures what it is presumed to measure. This is but one step better than the definition that states that a test is valid if it measures the truth. In this chapter it will be held that validity is a highly relative concept. If the question is asked about any particular test, "Is this test valid?" the answer should be in the form of another question, "Is it valid *for what?*" Furthermore, just as we found in the preceding chapter that we cannot, strictly speaking, state any figure as representing *the* reliability of a test, so we cannot give a single number to indicate *the* validity of a test.

There was a time, unfortunately still not entirely past, when each test was supposed to measure some underlying variable that went by a label. It was a test of intelligence, of introversion, or of neurotic tendency. Those concepts, because of the fixed labels, were supposed to be qualitatively fixed, known, and defined attributes. In order to be valid, tests going by those names were expected to correlate highly with older, generally accepted criteria of those supposed entities. For example, new tests were "validated" by demonstrating a strong correlation with the Stanford Revision of the Binet Test or with Laird's Test C2 or with Woodworth's Inventory.

Factorial Validity.—Now that these popular areas of personality have been shown to lack real unity and unanimity of reference,[1] we are properly more wary of attaching such labels to tests. If we regard intelligence as having been broken down into a collection of functional unities, called *primary abilities* for convenience, we find that the question of what is a valid intelligence test becomes meaningless. The primary abilities, on the other hand, have been arrived at by means of well-defined steps and can be verified by one who repeats those steps. If one acquiesces in the procedures by which those functional unities are discovered, he has no choice, if he still is concerned about the validity of tests, but to ask whether test *A* is a valid one for this primary ability or that one.

The validity of a test as a measure of one of these factors is indicated by its correlation with the factor, which is its *factor loading*.[2] It is recognized by those who adopt the factor-analysis approach that scarcely any test is an unadulterated measure of any primary ability or trait. Not only is it diluted by errors of measurement, as we saw in the discussion of reliability, but it is also adulterated with variances in other primary abilities or traits. This situation is overcome to some extent by a careful combining of tests, an exacting procedure that we cannot go into here. It is the author's belief that the best answer to the question, "What does this test measure?" is in the form of a list of the primary factors with which it correlates and their proportions of variance in the test.[3] This kind of validity may be called *factorial validity*. This idea will be explained

[1] See in particular Thurstone, L. L. Primary mental abilities. *Psychometr. Monogr.*, 1939, **1**; Guilford, J. P., and Guilford, R. B. Personality factors D, R, T, and A. *J. abnorm. soc. Psychol.*, 1939, **34**, 21–36; and Mosier, C. I. A factor analysis of certain neurotic tendencies. *Psychom.*, 1937, **2**, 263–286.

[2] For a brief discussion of factor theory and methods, see Guilford, J. P. Psychometric methods. New York: McGraw-Hill, 1936. Ch. XIV. In the discussion here, uncorrelated factors are assumed.

[3] Guilford, J. P. Factor analysis in a test-development program. *Psychol. Rev.*, 1948, **55**, 79–94.

more fully and it will be shown that it is basic to the understanding of other kinds of validity and of many phenomena of correlation in general.

Practical Validity.—The vocational counselor and the vocational selector have faced a different kind of problem when they inquire about validity of tests. They are concerned about predicting outcomes in specified tasks and situations—clerical ability, scholastic ability, salesmanship, and the like. A test is a valid one for clerical aptitude if its scores correlate highly with later clerical proficiency. Another test is a valid one for aptitude in selling, because it correlates highly with later proficiency in selling. From this point of view, any test is valid for any sphere of behavior if it enables us to predict within that sphere, regardless of the name of the test or the supposed fundamental abilities that it measures. A test designed to predict the success of student aviators may prove also to be a valid test of scholastic aptitude in engineering or of aptitude for a military career in general. From the practical standpoint, the validity of a test is its forecasting efficiency in predicting any measurable aspect of daily living.

Criteria for Validity.—One of the most difficult of all aspects of the validity problem is that of obtaining adequate criteria of what we are measuring. The factor-analysis approach has a fairly good solution when it is primary traits or abilities that we wish to measure. If two or more tests or items are combined to predict the factor, the validity coefficient is the multiple correlation between the tests and the factor. But practical criteria are most in demand and are most difficult to obtain and to measure adequately. An example of this is the criterion of scholastic achievement.

It has often been assumed that scholastic achievement, like intelligence, is a unitary attribute of each individual. But this is far from the truth. Although there is generally a positive correlation between achievement in different school subjects, there is sufficient disagreement to permit an individual to receive marks all the way from A to F in different subjects. It is best procedure, therefore, to examine the validity of each test used for guidance purposes in connection with *every* school subject taken by itself. Where a certain test of ability may possess only a moderate or low correlation with averages of school marks, it may correlate very high with specific courses. The writer has data showing correlations all the way from .37 to .74 between the Ohio State Psychological Examination, Form 20, and marks in freshman courses at a certain university. The point is that success in any sphere of life is ordinarily highly complex and is determined by many psychological factors in the individuals competing, rather than one or a few. If we measure success in a complex activity by singling out as criteria one or more of its aspects and measuring them, we are checking upon the validity of the test or tests for predicting those chosen aspects.

We should not identify those few aspects with the entire activity. We should, of course, attempt to single out the most significant aspects as criteria. Too often some inconsequential aspects are chosen because of their ready observability and measurability.

Having chosen the measurable variables of success in the area predicted, we have the problems of securing dependable measurements and perhaps of combining and weighting them in the wisest manner. With reference to measures of achievement, again, it should be emphasized that school marks as ordinarily assigned by teachers are rather poor metric material. Variations in meaning and standards from teacher to teacher and from course to course are notorious. Most marks are neither very reliable nor very valid indicators of achievement. The best measures of achievement in most courses are those obtained directly from good, comprehensive examinations of the objectively scored type. Marks otherwise obtained often have reliabilities in the range from .60 to .80, and their validities are unknown. When we attempt to find the predictive value of a psychological test, therefore, shall we reject tests that fail to correlate highly with such fallible criteria? We can allow for the unreliability of criteria statistically when we know a coefficient of reliability for them. We cannot so easily know or allow for lack of *validity* of criteria, though we can make allowances, knowing the kind of criteria we have.

A Brief Introduction to Factor Theory

Because so many of the facts of validity are explainable on the basis of factor theory, it is desirable for us to examine the basic features of factor theory in order to gain a better grasp of the problems and methods involved. There is not space here to describe the procedures for making a factor analysis of tests. These statistical procedures when described sufficiently for general use would take up a small volume in themselves.[1]

Basic Assumptions in Factor Theory.—It is best to begin with basic assumptions, two of which will give us the foundation we need for the logic of validity.

Assumption I: The total variance of a test may be regarded as the sum of three kinds of independent component variances: (1) that contributed by one or more common factors, *common* because they appear in more than one test; (2) that unique to the test itself and possibly to its equivalent forms; and (3) error variance. We are now ready to break up what was called *true* variance in the preceding chapter into component variances. Both the common-factor variances and the specific variance

[1] The most complete source of information on factor analysis is Thurstone, L. L. *Multiple factor analysis.* Chicago: University Press, 1947.

in a test contribute to its internal-consistency reliability, and to its equivalent-forms reliability. It is not necessary to assume that the common-factor and specific-factor variances are all independent or uncorrelated. To do so relieves us of having to deal with covariance terms and thus simplifies the picture. What follows would be just as true, in general, if we did not add this specification to the assumption.

Assumption I may be stated in the form of an equation:

$$\sigma^2_t = \sigma^2_a + \sigma^2_b + \cdots + \sigma^2_n + \sigma^2_s + \sigma^2_e$$

(Sum of independent variances in scores on a test) (18.1)

where σ^2_t = total variance of a test.
$\sigma^2_a, \sigma^2_b, \ldots, \sigma^2_n$ = variances in factors $A, B, \ldots,$ to N, respectively.
σ^2_s = variance specific to this test.
σ^2_e = error variance.

If we now divide equation (18.1) through by σ^2_t, we have

$$\frac{\sigma^2_t}{\sigma^2_t} = \frac{\sigma^2_a}{\sigma^2_t} + \frac{\sigma^2_b}{\sigma^2_t} + \cdots + \frac{\sigma^2_n}{\sigma^2_t} + \frac{\sigma^2_s}{\sigma^2_t} + \frac{\sigma^2_e}{\sigma^2_t} = 1.00 \qquad (18.2)$$

Substituting new symbols for these fractions, which are proportions, we have

$$1.00 = a^2_x + b^2_x + \cdots + n^2_x + s^2_x + e^2_x$$

(Proportions of factor variances in a test) (18.3)

where $a^2_x, b^2_x, \ldots, n^2_x$ = proportions of total variance contributed to test X by factors $A, B, \ldots, N$, respectively.
s^2_x = proportion of specific variance in test X.
e^2_x = proportion of error variance in text X.

In the same notation, the reliability of test X can be written as

$$r_{tt} = 1 - e^2_x = a^2_x + b^2_x + \cdots + n^2_x + s^2_x$$

(Reliability as a sum of proportions of nonerror variance) (18.4)

This equation will be useful in discussions of the relation of validity to reliability later on.

Communality.—A new concept that should be pointed out here, although we will not have occasion to do much with it in a practical way in this chapter, is known as the *communality* of a test. The communality of a test is the sum of the proportions of common-factor variances. In equation form,

$$h^2_x = a^2_x + b^2_x + \cdots + n^2_x \qquad \text{(Communality of a test)} \qquad (18.5)$$

The communality of a test contains all the nonerror variance except the specific variance. Communality is what gives any test the chances of correlating with other tests and with practical criteria. If there were no communality in a test it could be quite reliable and still not correlate with anything else. On the other hand, a test could have relatively low reliability, and yet if all of its nonerror variance were in common with variance in other variables, its correlations with other things could be rather substantial, hence its validity could be good.

As an example, let us consider three tests and a practical criterion. Five common factors are represented in these four variables. In Table 18.1 we have listed the proportions of common-factor, specific, and error variance for each variable. Test 1 has 36 per cent of its variance

TABLE 18.1.—PROPORTIONS OF COMMON-FACTOR, SPECIFIC, AND ERROR VARIANCE IN THREE TESTS AND A PRACTICAL CRITERION OF PROFICIENCY

Variable	Common factors					Specific S	Error E	Communality h^2	Reliability r_{xx}
	A	B	C	D	F				
Test 1	.36	.00	.36	.00	.00	.10	.18	.72	.82
Test 2	.16	.00	.12	.00	.64	.00	.00	.92	.92
Test 3	.00	.49	.00	.25	.00	.09	.17	.74	.83
Criterion J	.16	.09	.16	.25	.00	.14	.20	.66	.80

accounted for by factor A, and 36 per cent by factor C. The sum of these two components equals 72 per cent, which represents the communality of this test. Add the 10 per cent specific variance, and we have 82 per cent, which represents the test's true variance and a reliability of .82. The remaining 18 per cent is error variance. The other tests and criterion J can be interpreted in a similar manner. Fig. 18.1 shows the component variances for these same four variables, each as a segment of a bar diagram.

Factor Loadings.—The proportion of a total variance contributed by one component may be regarded as a coefficient of determination of the total by the part. The square root of each proportion of variance contributed by a common factor may therefore be regarded as the correlation between the total variable and the factor. These square roots are correlation coefficients and are known as *factor loadings* or *factor saturations.* For the three tests and criterion J, the common-factor loadings are given in Table 18.2. Test 2 correlates .40 with factor A, .35 with factor C, and .80 with factor F. Factor F has no correlations with other variables in this list, but in order to be regarded as a common factor it must have

some correlation with other variables not in this list. The square roots of specific variance are not listed because it is not certain what the specific variances represent. A certain specific variance may indeed be unique

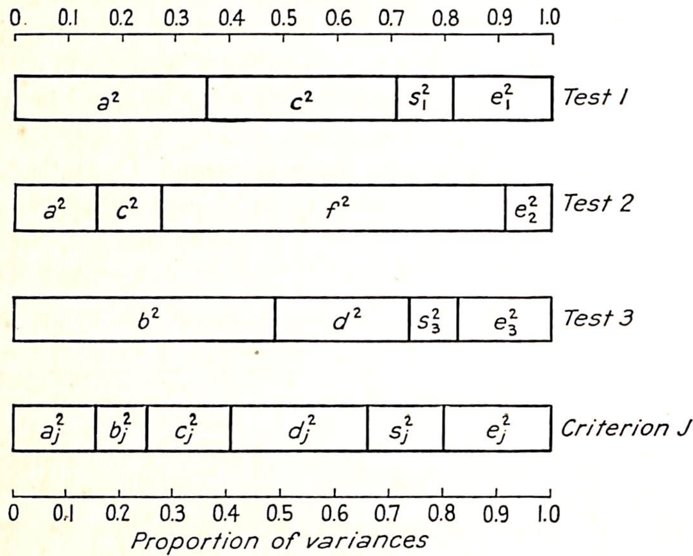


FIG. 18.1.—Proportions of common-factor, specific, and error variance in three hypothetical tests and a criterion.

to its own test, but it may be a composite of some kind, in which case each component of the specific variance would have its own correlation with the total. On the other hand, some specific variances might turn out on later analyses to be one or more unrecognized common-factor variances. Certain tests have been known to lack any specific variance at all; the entire true variance being composed of common-factor components and the communality equaling the reliability of the test.

TABLE 18.2.—FACTOR LOADINGS (CORRELATIONS OF COMMON FACTORS WITH EXPERIMENTAL VARIABLES) FOR THE THREE TESTS AND A CRITERION

Variables	Common factors				
	A	*B*	*C*	*D*	*F*
Test 1	.60	.00	.60	.00	.00
Test 2	.40	.00	.35	.00	.80
Test 3	.00	.70	.00	.50	.00
Criterion *J*	.40	.30	.40	.50	.00

Assumption II: The second major assumption of factor analysis is that the correlation between two experimental variables (such as tests and criteria) is equal to the sum of the cross products of their common-factor loadings. In equation form,

$$r_{jx} = a_j a_x + b_j b_x + \cdots + n_j n_x \quad \text{(A correlation as a sum of factor-loading products)} \qquad (18.6)$$

where a_j and a_x = loadings of factor A in criterion J and test X.
b_j and b_x = loadings of factor B in criterion J and test X, etc.

How Factor Theory Explains Practical Validity.—Applied to the loadings given in Table 18.2, the correlation between tests 1 and 2 would be

$$r_{12} = (.6)(.4) + (.0)(.0) + (.6)(.35) + (.0)(.0) + (.0)(.8) = .45$$

The correlation between test 1 and criterion J (its validity for predicting criterion J) would be

$$r_{j1} = (.4)(.6) + (.3)(.0) + (.4)(.6) + (.5)(.0) + (.0)(.0) = .48$$

The other intercorrelations and validity coefficients found in similar manner are listed in Table 18.3. In experimental practice we do not know the factor loadings first and derive from them the intercorrelations; we know the intercorrelations and by factor analysis arrive at the factor loadings. We have assumed that the factor loadings are known here for the sake of illustration.

Examination of the three validity coefficients in Table 18.3 shows that they are .48, .30, and .46, for tests 1, 2, and 3, respectively. The three validity coefficients are represented graphically in Fig. 18.2. The reasons

TABLE 18.3.—INTERCORRELATIONS OF TESTS AND CRITERION J DERIVED FROM THEIR COMMON-FACTOR LOADINGS

Variables	Tests			Criterion J
	1	2	3	
Test 1	—	.45	.00	.48
Test 2	.45	—	.00	.30
Test 3	.00	.00	—	.46
Criterion J	.48	.30	.46	—

for the validity of tests 1 and 2 are the same; their common ground with the criterion is in factors A and C. The reason test 3 is valid, however, is totally different from this. Test 3 is valid because of having in common

with the criterion factors B and D. Test 2 has the lowest validity for predicting criterion J, but its unusually large loading in factor F offers strong possibilities for its validity in predicting some other criterion that has a substantial loading in Factor F.

How Factor Theory Explains Multiple Correlation Principles.—The multiple correlations of some of these tests and criterion J can be nicely explained by the various factor loadings. The multiple correlation $R_{j.12} = .49$, which is only .01 higher than the correlation r_{j1}. Adding test 2 in a battery to test 1 to predict J is of little value because both bring to the composite a coverage of the same common factors in J. The multiple $R_{j.13}$, however, is equal to .66. Adding test 3 to test 1 to make a

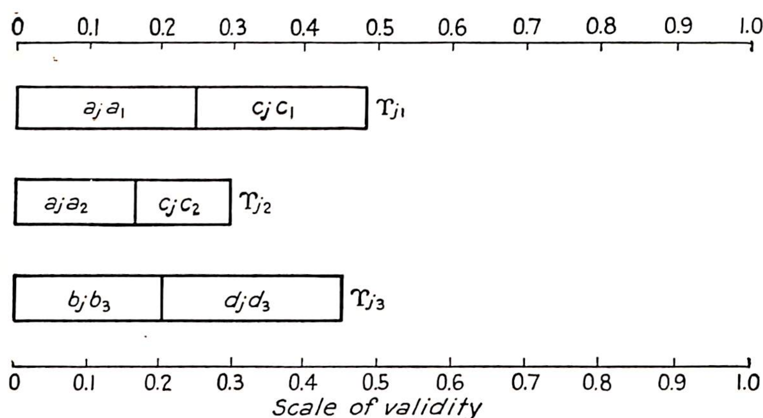


FIG. 18.2.—Segments of three intercorrelations of tests and a criterion that are contributed by different common factors.

joint prediction of J is very effective because the two tests cover totally different components in J. The multiple $R_{j.23}$ is less than $R_{j.13}$, being .55. The reason for this is that test 2 does not cover factors A and C nearly so well as does test 1.

We might well raise the question at this point as to whether tests 1 and 3, optimally weighted, with their multiple R of .66, have yielded the maximum amount of validity possible for a weighted composite that contains factors A, B, C, and D. Reference to equation (18.6) will show that the correlations r_{j1} and r_{j3} could have been higher if the tests' factor loadings a_1, c_1, b_3, and d_3 had been larger. The only limits to those factor loadings would be that the communalities should not exceed 1.0. This, however, is not the whole story. We could make those loadings as large as the communalities would allow and they would still not yield the maximal correlation with criterion J unless they were in the right proportions. The right proportions would have to take into consideration

the proportions of loadings a_j, b_j, c_j, and d_j in the criterion. With sufficient loadings of the four factors in the tests and with proper weightings, the maximum validity for the composite in predicting criterion J would be equal to the square root of the communality of that criterion. The square root of .66 is .81. This principle is reminiscent of the one mentioned in the last chapter regarding the index of reliability, which is the square root of the reliability coefficient. It gives the maximum possible correlation of anything with the variable in question. In this statement, however, is latent the assumption that all the true variance is common-factor variance; that $h^2 = r_{tt}$.

It is doubtful whether tests 1 and 3 could ever be weighted appropriately to yield a validity for their composite equal to the maximum .81 with criterion J, even though their common-factor loadings were as large as possible. The reason is that factors A and C are tied together in the same test and factors B and D are tied together in the other test. Since factors A and C have equal loadings in criterion J and also in test 1, as long as they keep the same ratio in test 1 they would be properly weighted in a regression equation. This is merely a coincidence in this particular problem. Factors B and D, however, are weighted in reverse order in test 3 and criterion J. For optimal prediction of J, the loading d_3 should be greater than the loading b_3, to correspond with the fact that the loading d_j is greater than b_j. If we had loadings b_3 and d_3 in proportion to the loadings b_j and d_j and also 50 per cent larger (just as a_1 and c_1 are 50 per cent larger than a_j and c_j), they would be .45 and .75, respectively. These would yield (by equation 18.6) an r_{j3} equal to .51 (where it was .46) and a multiple R of .70 (where it was .66). The moral of this is that for the freedom to weight each factor in a composite as it should be weighted to get the maximal prediction of a criterion, it is best to use unique, or univocal, tests; that is, each test with but one common factor. In practice, a regression weight has to be applied to the test as a whole and all factors in it are weighted the same, in so far as external weights are applied.

We have just seen that increasing the practical validity of a composite depends upon large factor loadings for factors represented in the criterion and an *optimal weighting of the individual factors*. There is another important way of increasing the validity of a composite, and that is to bring in a new test that covers a common factor in the criterion that is not already covered. Criterion J was reported to have 14 per cent of its variance devoted to specific sources. It is possible that this portion of the variance in J is really contributed by an unknown common factor. Further experimental work might lead to an identification of it as stemming from

one or more common factors. Suppose that it were found to belong to one additional factor *G*. To contribute .14 to the total variance, the loading g_j would be about .37. With an additional test to measure this factor in the composite, the multiple *R* could be increased materially. On the whole, there is much more to be gained in increasing *R* by discovery or identification of new factors than there is by increasing loadings for already known factors. With a large number of factors in a criterion, sizes of loadings will have to be small in order to stay within the limit of its communality, and their multipliers (loadings in the tests) can be correspondingly small, so as to produce a validity coefficient within the limit of the square root of that communality.

Conditions upon Which Validity Depends

Relation of Validity to Reliability.—It has been a common belief that the practical validity of a test, other things being equal, is directly proportional to its reliability—the more reliable a test, the more valid it is. There is much in the application of factor theory to support this idea, as we can see by reference to previous paragraphs. The greater the error variance in a test, the less room there is for common-factor variance, and common-factor variance is the source of validity. If we make a test more reliable and in so doing we increase variances in common factors, the possibilities for validity should be increased accordingly.

When Validity and Reliability Are Independent.—There are important exceptions to this relationship between validity and reliability. If a test is heterogeneous, we might have a very low internal-consistency reliability and yet a high practical validity. If a test is homogeneous, we might increase its reliability without affecting its validity. The increased reliability might mean added variance in a common factor that has no relation to the criterion. For example, a test measuring visualization is known to have validity for the selection of pilots. We might increase the reliability of this test by making it more difficult, thereby adding reasoning variance. Reasoning variance has no correlation with the pilot criterion, consequently no improvement in pilot validity would follow for this test. The added common-factor variance in a test will increase the practical validity of a test only when that new type of variance is also present in the criterion. If there were no valid variance in a test to begin with, no amount of increased reliability will give it validity unless the added variance is related to the criterion.

Goals of Validity and Reliability Often Incompatible.—When we seek to make a single test both highly reliable (internally) and also highly valid, we are often working at cross purposes. The two goals are incompatible

in many respects. In aiming for one goal we are likely to defeat our efforts toward the other.

Maximal reliability requires high intercorrelation among items; maximal validity requires low intercorrelations. Maximal reliability requires items of equal difficulty; maximal validity requires items differing in difficulty. This point needs some explanation. Tucker has demonstrated this fact mathematically, but there is a simpler, common-sense rationale.[1] A range of difficulty is necessary, of course, in order to obtain graded measures of individuals. It was shown in Ch. 17 how with perfect intercorrelation of items (which could occur with ϕ coefficients only when items are of exactly equal difficulty) there were only two scores—perfect scores and zeros. For spacing individuals in fine enough gradations for measurement purposes it is necessary to have a continuous distribution, not a U-shaped one. It would be ideal, for fine measurements, to space items, each discriminating well between all those above a certain point on the scale and those below, rather evenly all along the range of ability in the population. With such spacings, intercorrelations could not be perfect and some would indeed be very low.

There must be some compromising of aims; both reliability and validity cannot be maximal. Fortunately, the kind of moderate item intercorrelations usually obtained for well-constructed items are of the size that, according to Tucker's conclusions, will yield good validities. They will also yield satisfactory reliabilities, but those reliabilities will not often be above .90. To be more specific, the item-test correlations for well-constructed items range between .30 and .80, which means item intercorrelations probably between .10 and .60. Items within these ranges of correlation should provide tests of both satisfactory reliability and validity. There is probably better reason for going below these limits than above them in constructing items. To do so would probably err on the side of validity, which, after all, is the more important.

Homogeneous Tests; Heterogeneous Batteries.—The relation of heterogeneity to validity deserves more attention. One way to make a test more valid is to make it more heterogeneous. In factorial language this means adding new factors. If we succeeded in getting into the scores of the single test all the factors that are also in the practical criterion, and if we weighted them properly, we could achieve maximal accuracy of predictions from the single test.

Recall, in this connection, the principles of the multiple regression equation. Maximal multiple correlation is achieved by minimizing the inter-

[1] Tucker, L. R. Maximum validity of a test with equivalent items. *Psychom.*, 1946, **11**, 1–13.

correlations of the independent variables. If we apply this to test items, as separate variables, the principle still holds. The ideal test, from this point of view, would be one in which each item measured a different factor (and measured it accurately). This would mean a test of low internal reliability. It would also mean a test, which, though correlating well with the criterion, would make very crude discriminations for each factor. Each item would differentiate only two categories—those who pass it and those who fail it—for each trait measured. If we brought in a number of items to measure each factor, with differences in difficulty to overcome this defect, we would have virtually a battery of tests within a single test.

The solution to the incompatability of goals of reliability and validity is precisely as just suggested; to use a battery of tests rather than single tests. Reliability should be the goal emphasized for each test; validity the goal emphasized for the battery. Even in the single test some reliability should be sacrificed for the sake of well-graded measurements. It is strongly urged that, if possible, each test be designed to measure one common factor. It should be univocal; its contribution unique. In this way minimal intercorrelations of tests is assured, which satisfies one of the major principles in multiple regression. It was also shown that when tests are univocal the various factors can be weighted in the best way to make each prediction. The univocal test will correlate less with a practical criterion than will a heterogeneous test, but what we lose in validity for the single test will be more than made up by forming batteries which cover the factors to be predicted and in a more manageable manner. For the sake of meaningful profiles also, a battery of univocal tests has no equal.

Reliabilities and Test Batteries.—If a composite score from a battery is to be used and not part scores from the components, as in a profile, it is likely that there is not much to be gained by achieving reliabilities for single tests higher than .60 or by having tests longer than 30 items each.[1] The reliability of the composite score of independent tests will be approximately a weighted average of the reliabilities of the components.[2] This means that if the components have a generally low reliability, in such a battery the reliability of the composite will be low. This need not be disturbing, providing the validity of the composite is high. To the extent that the components are intercorrelated, the reliability of the composite will exceed the average reliability of the components. In general, if there

[1] Dailey, J. T. Determination of optimal test reliability in a battery of aptitude tests. Technical memorandum No. 10, Lackland Air Force Base, 1948.

[2] Mosier, C. I. On the reliability of a weighted composite. *Psychom.*, 1943, **8**, 161–168.

is a choice between lengthening of tests in a battery to make them more reliable and adding more tests of different kinds that contribute unique valid variances, the decision should certainly go to the second alternative. If part scores are to be used separately, however, attention must also be given to reliability of components.

Discrimination Values of Items.—Some of the points just discussed may be made a little clearer if we approach the item theory from a still different aspect. Fig. 18.3 is used to illustrate this approach. Imagine a scale of ability or of any other trait that we attempt to measure by means of a test.

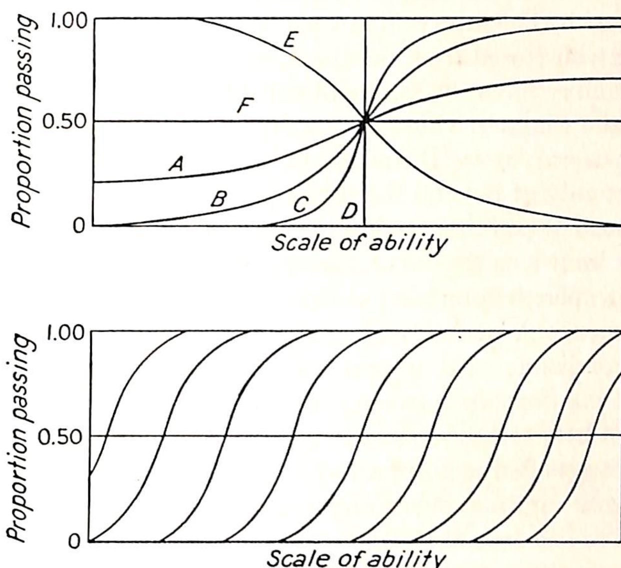


Fig. 18.3.—Proportion passing an item (responding correctly) as a function of ability level on the scale of the kind of ability (or weighted combination of abilities) required to pass the item.

We want each item to correlate with that variable; to predict the status of individuals with respect to the variable; to discriminate between individuals.

Suppose we already know the positions of large numbers of individuals on this scale. We apply to them an item that we will call item *C*. The item is of median difficulty, for of the entire group 50 per cent respond in the acceptable manner and 50 per cent do not. According to the requirements of good reliability, this knowledge about the difficulty of item *C* is promising, but not sufficient evidence that the item would contribute to a reliable test. We do not yet know whether it is at all related to the variable we want to measure. It could be of median difficulty and still be

uncorrelated with other items in the test. Let us subdivide the large sample into subsamples grouped in class intervals as if for known values along the scale. We are now interested in seeing whether those groups higher on the scale have any greater probability of passing the item than those lower on the scale. Theory states, and experimental evidence supports the idea, that the increase in the probability of passing the item follows the normal cumulative frequency curve. The regression of proportions passing the item upon ability is the S-shaped or ogive form. For item C, not very far below average ability we find a point below which none pass the item. Above a point just as far above the mean we find that all pass the item. The interval between is sometimes called the *transition zone*, a concept borrowed from psychophysics.[1]

Other items may have the same difficulty level as item C, but like items B and A in the diagram (Fig. 18.3) they have different degrees of discriminating power. Both B and A have much wider transition zones (they both actually go beyond the range of the given horizontal scale) and their curves have slopes that are less steep than that for C. The steepness of the slope is known as the curve's *precision*. The term applies well here because the steeper the precision of the curve, the greater is the precision of discrimination. A perfectly discriminating item is D, whose slope is infinite. A nondiscriminating item is F, whose slope is zero. There is a mathematical relationship between the precision of an ogive like these and the correlation between the item and a good measure of the trait.[2] Item E would have a negative correlation with the variable to be measured. This would be an unusual event and would probably mean that the item was keyed wrong in scoring. Items like D would seem to be ideal; they are perfectly discriminating. But it can be seen how only one such item used alone would be almost futile, for it discriminates at only one point.

The second diagram is more realistic and yet pictures a somewhat ideal situation. It shows a series of items equally spaced as to difficulty and all with excellent discriminating power. With the extensive range of difficulty level, there could not be as high internal reliability as some might desire. But the possibility of accurately grading individuals on a continuous scale, and hence the possibility of good validities, is greater because of that dispersion. To appreciate the full value of the items that depart from medium difficulty, one would need either to use a tetrachoric r in correlating item with total score or to make allowance for the effect of divergencies in difficulty upon the phi coefficient.

[1] Woodworth, R. S. Experimental psychology. New York: Holt, 1938. P. 401.

[2] For proof of this see Richardson, M. W. Relation between the difficulty and the differential validity of a test. *Psychom.*, 1936, **1**, 33–49.

Validity and the Length of Test.—Since the homogeneous lengthening of a test increases its reliability, in accordance with the Spearman-Brown formula, it will also increase its validity. If the change in length is by some ratio n (the new length divided by the old) the new validity of the test is estimated by the formula.

$$r_{y(nx)} = \frac{r_{yx}}{\sqrt{\frac{1 - r_{xx}}{n} + r_{xx}}} \quad \text{(Validity of a homogeneous test increased } n \text{ times)} \qquad (18.7)$$

where r_{yx} = validity coefficient for predicting criterion Y from test X.
r_{xx} = reliability of test X.

A certain line-drawing test developed to predict creative abilities of students in a course in designing had a reliability of .57 and a correlation with teacher's ratings of .65.[1] If this test were made twice as long, what validity could be expected? Applying formula (18.7),

$$r_{y(2x)} = \frac{.65}{\sqrt{\frac{1 - .57}{2} + .57}} = \frac{.65}{.886} = .73$$

It would thus definitely pay to make this test longer and more reliable in order to improve its validity.

If we wanted to know how much homogeneous lengthening is needed in order to achieve a desired level of validity, we could do this by solving formula (18.7) for n, which gives

$$n = \frac{1 - r_{xx}}{\frac{r^2_{yx}}{r^2_{y(nx)}} - r_{xx}} \quad \text{(Ratio of new length of test for a required validity)} \qquad (18.8)$$

where the symbols are as defined for formula (18.7).

If we wanted a validity of .80 for the line-drawing test, the revised length would have to be

$$n = \frac{1 - .57}{\frac{.4225}{.64} - .57} = \frac{.43}{.0901} = 4.8$$

[1] Guilford, J. P., and Guilford, R. B. A prognostic test for students in design. *J. appl. Psychol.*, 1931, **15**, 335–345.

Whether it would be practical to devote nearly five times as much effort to this test is a question of policy that goes beyond statistical answers.

Relation of Validity Coefficients to Errors of Measurement.—When two fallible measures are correlated, the errors of measurement, if uncorrelated among themselves, always serve to lower the coefficient of correlation as compared with what it would have been had the two measures been perfectly reliable. We say that the degree of correlation has been attenuated. If we want to know what the correlation would have been if the two variables were perfectly measured, we must resort to the *correction for attenuation*, for which we have a formula

$$r_{\infty\omega} = \frac{r_{xy}}{\sqrt{r_{xx}r_{yy}}} \quad \text{(An intercorrelation corrected for attenuation)} \quad (18.9)$$

where r_{xx} and r_{yy} = reliability coefficients of the two tests.

The correlation obtained between a figure-classification test and a form-perception test was only .36. The reliability coefficients for the two tests were .60 and .94, respectively. Applying formula (18.9),

$$r_{\omega\infty} = \frac{.36}{\sqrt{(.60)(.94)}} = \frac{.36}{.751} = .48$$

We should therefore expect the correlation between true scores in these two tests to be .48 rather than the obtained one of .36.

In general, when making this correction for attenuation in both fallible tests, if we are dealing with two forms of the same test for purposes of finding reliability, there is a possibility of determining four intercorrelations between the two tests, *i.e.*, each form of the one correlated with the two forms of the other. In this case, it is well to use all the information we can get concerning the intercorrelation of the two tests by computing the four coefficients and using their arithmetic mean as a better estimate of the numerator of the fraction in formula (18.9).

Factorial Explanation of Attenuation and Its Correction.—It may not be clear to the reader why errors of measurement always lower intercorrelations, and why, when the corrective formula is applied, correlations should not be perfect. The answers to both of these questions can best be given by reference to factor theory.

Consider test 1 and criterion J, of the illustration used above when factor theory was introduced. Error variance made up 18 per cent of the

total variance of test 1 and 20 per cent of criterion J. Let us suppose that we could rid each variable of all errors of measurement; all error variance. In doing so, let us further suppose that the remaining true variance is expanded with all of its components in proportion to their original amounts. Figure 18.4 demonstrates what happens when the error components are "squeezed out" of variables and the true-variance components expand to take their places. Variances that were .36 and .36 in factors A and C in test 1 before correction become .439 and .439 after correction. The new factor loadings are .663 in each factor. In the criterion the corre-

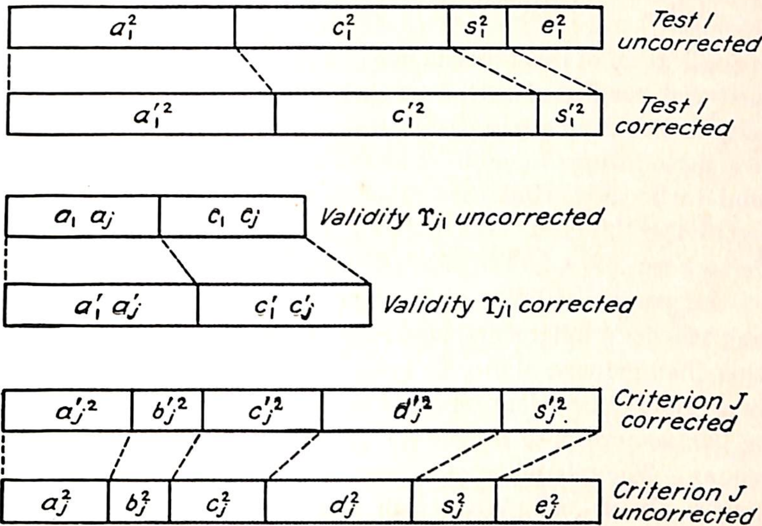


FIG. 18.4.—Proportions of variance in a test and a criterion after correction for attenuation (elimination of error variance statistically), also the contribution of factors to the validity coefficient before and after correction.

sponding loadings become .447. By equation (18.6), the new correlation r_{j1} becomes .59, whereas it was .48. The use of formula (18.9) applied to the original r_{j1} gives

$$r_{\infty\omega} = \frac{.48}{\sqrt{(.82)(.80)}} = \frac{.48}{.81} = .59$$

The change in validity from .48 to .59 is shown graphically in Fig. 18.4.

Correction for Attenuation in the Criterion Only.—The preceding device has limited application except in theoretical problems. In practice, we are compelled to deal with fallible tests. If the tests from which we wish

to predict something else are not perfect, that fact must be faced, and our predictions are reduced in accuracy accordingly. But we should hardly expect to be asked to overlook the fallibility of the criterion we are trying to predict. If it measures success inaccurately, this lack of accuracy should not be permitted to make it appear that the test is less valid than it really is. It is becoming more customary, therefore, to correct validity coefficients for attenuation in the criterion measurements but not in the test scores. This one-sided correction is made by the formula

$$r_{\infty x} = \frac{r_{xy}}{\sqrt{r_{yy}}} \quad \text{(Validity coefficient corrected for attenuation in the criterion only)} \qquad (18.10)$$

As an application of this formula, we cite the line-drawing test previously mentioned that correlated with a teacher's rank-order judgments of creative ability in her students in design to the extent of .65. The reliability of the teacher's ratings (combined from two rank orders a month apart) was found to be .82. Had the teacher's ratings been perfectly reliable measures of the thing she was judging, the correlation with test scores would have been $.65/\sqrt{.82} = .72$. The correlation of .72 is accordingly taken as the genuine validity of the test, unless we are concerned about predicting teacher's judgments, contaminated by flaws as they obviously are, rather than genuine ability as evidenced by those ratings.

Many a validity coefficient reported in the literature is of very uncertain meaning because errors of measurement in the criterion were not taken into account. The reliability of ratings, even of the better ones, is characteristically about .60. With such criteria, validity coefficients are about 25 per cent underestimated. Too often the reliability problem of a criterion is entirely ignored. The writer has known of purported criteria of a performance kind (bombing errors of bombardiers in training) which at best had reliabilities of only approximately .30. What is even more important, but incidental to the discussion here, is the *validity* of the criterion. Any investigator who hopes to develop successful selective instruments is often beaten before he starts, if he does not first insure reliable and valid criteria, or if he does not estimate these features and make allowances for them.

Limitations to the Use of Correction for Attenuation.—The correction of a correlation for attenuation requires that we have a rather accurate estimate of reliability for each variable that enters into the situation. If either r_{yy} or r_{xx} is underestimated, the corrected r_{yx} will be overestimated. If either reliability index is overestimated, the corrected r_{yx} will be underestimated. It is probably best, if one wishes to be on the conservative side, that, if anything, a reliability estimate should be too large when used

for this purpose. On the other hand, it is likely that most estimates of internal-consistency reliability are too low, which is in the wrong direction for conservatism.

There is also the question as to which of the three main types of reliability coefficient is desirable in correcting for attenuation. There are proponents for the use of each type in this connection. It is best to decide what kind of errors of measurement should be ruled out in the particular situation or particular use of r_{tt}. Once this decision is made, the type of reliability will be selected accordingly, since it was shown in the preceding chapter that each type emphasizes certain sources of variance as error. The tendency of underestimation of r_{tt} by internal-consistency methods is against their use when there is a reasonably good alternative.

The Index of Forecasting Efficiency with a True Criterion.—An index of forecasting efficiency (see Ch. 15) could be computed directly from $r_{\infty x}$ to denote the improvement in predicting the true criterion variable on the basis of knowledge of test scores over prediction without that knowledge. This statistic can be calculated directly from the known r's, however, without first finding $r_{\infty x}$, by use of the formula[1]

$$E_{\infty x} = 100\left(1 - \sqrt{1 - \frac{r^2_{yx}}{r_{yy}}}\right) \quad \text{(Index of forecasting efficiency of a true criterion)} \qquad (18.11)$$

Standard Error of the Estimate of a True Criterion.—Taking the correlation between our fallible scores and an infallible or true criterion as the coefficient of validity, we shall also have smaller errors of prediction than if we tried to predict fallible criterion measurements. We could substitute $r_{\infty x}$ in the usual formula for finding the standard error of the estimate from r, but the σ_{yx} (which now becomes $\sigma_{\infty x}$) can be calculated directly from the original correlations by the formula

$$\sigma_{\infty x} = \sigma_y \sqrt{r_{yy} - r^2_{yx}} \quad \text{(Standard error of estimate of a true criterion)} \qquad (18.12)$$

Level of Difficulty and Validity of a Test.—The validity of a test may be seriously influenced by its level of difficulty. In subjecting the Seashore Test of Pitch Discrimination to a factor analysis, the author discovered that at different levels of difficulty, three distinctly different abilities were being measured.[2] It is possible that among the easiest items, where differences between pairs of tones were 17, 23, and 30 cycles per second,

[1] Conrad, H. S., and Martin, G. B. The index of forecasting efficiency for the case of a "true" criteron. *J. exp. Educ.*, 1936, **4**, 231–244.

[2] Guilford, J. P. The difficulty of a test and its factor composition. *Psychom.*, 1941, **6**, 67–77.

individual differences in scores represented differences in attentiveness to an easy task, and errors were made because of lapses of attention. Some of the more difficult items, having differences of 1 to 5 cycles, were most heavily correlated with a second factor or ability. Items with differences from 3 to 12 cycles were most heavily correlated with a third factor or ability. The moral is that a total score limited to one of the three ranges would be most valid for that particular range. A total score based upon the total range of difficulty from 1 to 30 cycles would represent some kind of composite measurement.

It is likely that other tests, likewise, are altered in validity as they are easy or difficult for the group examined. Very easy tests may become measures of perceptual abilities or of motor speed, and more difficult tests of similar items may become measures of reasoning abilities of one kind or another.

Validity of Speed Tests and of Power Tests.—The problem of difficulty level of items is intrinsically related to the problem of time limit of a test. A test of very easy items would have to be given with strict timing in order to prevent serious negative skewing of distributions, if not to achieve any variance at all among scores. A moderately difficult test would have to allow ample time or there would be serious positive skewing and loss of variance among scores. It has often been assumed that for tests of moderate difficulty it does not matter much what the time limit is, so long as not too many examinees finish before the time is up. It is sometimes maintained that those who finish and do nothing until time is called could have been adding more points to their scores. The implicit belief in this position is that there is an interchangeability between level of difficulty an individual can master and the speed with which he can solve problems of equal difficulty. This hypothesis still lacks proof.

Thurstone has made a beginning toward rationalizing the problem of speed versus power in performing tasks.[1] His thesis is that power is best measured by allowing the examinee "infinite" time, which, in practice, would be all the time he will take. Speed is best measured by the number of tasks of equal difficulty performed in a given time. The two aspects of performance behave differently. Increased motivation cannot raise the individual's power, but it can increase his speed of performance. If all individuals have equal motivation, a speed score with items of sufficient difficulty level might still reflect individual differences in power, if accuracy is also weighted. It should also follow that the more the speed aspect enters into the variance of scores, the more individual differences in motivation might become apparent in the scores.

[1] Thurstone, L. L. Ability, motivation, and speed. *Psychom.*, 1937, **2,** 249–254.

A number of experimental studies have shown substantial, if not high, correlations between scores made on the same test under both power and speed conditions. The experiments were not always well controlled, and there have been exceptions in which low correlations were found. A better controlled study of Davidson and Carroll in which factor analyses were made of speed scores (rate of work without regard to accuracy), power scores (number of right answers when all items were attempted), and a mixed score (number of right answers when not all examinees attempted all items).[1] One finding was that speed scores, as defined, are practically independent of the power scores. Another was a speed-of-work factor, common to all the speed scores. Speed and power scores may measure to some degree some of the same factors, but to different extents. The results are very suggestive of the changes in factorial content of a test as its time limit is altered and as scoring emphasizes speed or accuracy.

As a general conclusion to this discussion on speed versus power, it is evident that there are some traits that are better measured by speed tests and there are others that are better measured by power tests. Still others may be best measured by means of some balance of speed and power conditions still to be determined by experiment. It will ordinarily pay the test maker to consider this problem in connection with any new test, and to reconsider it in connection with revisions of old ones.

Validity of Right and Wrong Responses.—Many tests are scored with a formula score in which the wrong responses are given a negative fractional weight and the right responses a weight of +1.

A Priori Scoring Formulas.—One of the reasons back of such scoring formulas is the a priori reasoning about chance success and the need for correcting for it. In a true-false test we have a two-alternative situation and the assumption is that when the examinee does not know an answer he will guess at random. When he guesses, his probability of getting the right answer is .5. When there are three alternatives, the theoretical proportion of right answers in guessing is .33; in a four-choice item the probability is .25, and so on. This has led to the stock scoring formula of the form

$$S = R - \frac{W}{n - 1} \quad \text{(A test score with a priori correction for guessing)} \qquad (18.13)$$

where R = number of right responses.
W = number of wrong responses.
n = number of alternative responses to each item.

[1] Davidson, W. M. and Carroll, J. B. Speed and level components in time-limit scores; a factor analysis. *Educ. & Psychol. Meas.*, 1945, **5**, 411–427.

In a true-false test this reduces to the familiar $R - W$. In a five-choice-item test it becomes $R - W/4$. Incidentally, a similar correction could be made by the general formula

$$S = R + \frac{O}{n} \quad \text{(Alternative scoring formula with correction for guessing)} \qquad (18.14)$$

where O = number of omissions.

It should be emphasized that neither of these formulas will tend to reduce the error variance introduced by guessing unless there are an appreciable number of omissions or failures to attempt items. If every examinee attempts all items, the correlation between R and W will be a perfect -1.0, which offers no freedom for improvement by scoring formula. The formula scores would then correlate $+1$ with R and the correction operation would be of no value. In a speed test, however, and in a power test in which the examinees voluntarily omit many items, such a scoring formula may help to eliminate some of the error variance and thus promote better reliability and validity.

If a scoring formula of this type is to be used in a test, and particularly if it is a power test, there should be explicit instructions to the examinees that there will be a deduction of a fraction of a point for each wrong answer (or a bonus of a fraction of a point for an omission). The second formula is naturally more palatable to examinees. But there are usually better scoring formulas than those based upon the a priori reasoning about guessing, as we shall see next. It might be pointed out, incidentally, that when examinees are ignorant of the answer to an item, their habits of taking tests are such that they do not choose among the alternatives entirely at random. Certain positions in a list of five responses may be favored by habits of reading or of attention. This is probably not sufficiently important in itself to overthrow the usefulness of "chance" scoring formulas. In the long run, if the position of the right answer is randomized, the correction may work well enough. More serious, however, is the fact that many test writers, in preparing four- or five-choice items, do not provide "misleads" or "distractors" that are equally attractive. It is easy, perhaps, to think of one good wrong answer to an item, but to think of more than one and to make all equally attractive is a trying art. Many a four- or five-choice item reduces virtually to a three- or two-choice item because of this fact. The a priori scoring formula as given above then undercorrects. We do not know by how much.

Empirical Weighting of Right and Wrong Answers.—When R and W scores are not too highly intercorrelated, and when there is a practical

criterion, it often pays to treat the two as if they were two different variables; as if they had arisen from two different tests. One then applies the multiple-regression procedures and derives optimal weights which will maximize the correlation of a weighted combination of R and W scores and the criterion. Since it is, as pointed out before, the *relative* sizes of the weights that are important and we do not care whether the formula scores have the same mean as the criterion or represent predictions in proper sizes, we can let the R score have a weight of +1 and find what weight the W score must then have. We would expect it to have a fractional negative weight, though it might differ markedly from the weight given by formula (18.13). For this purpose, Thurstone has given the following equation to determine the weight for the W score:[1]

$$v = \frac{\sigma_r(r_{cr}r_{wr} - r_{cw})}{\sigma_w(r_{cw}r_{wr} - r_{cr})} \quad \text{(Optimal weight for error scores when weight for rights scores is +1)} \qquad (18.15)$$

where the subscripts c, r, and w stand for criterion, rights, and wrongs scores, respectively. The correlation between these formula scores and the criterion is given by the usual multiple-R formula for three variables. In symbols that apply here,

$$R^2_{c.rw} = \frac{r^2_{cr} + r^2_{cw} - 2r_{cr}r_{cw}r_{wr}}{1 - r^2_{wr}} \quad \text{(Correlation of optimally weighted formula score with a criterion)} \qquad (18.16)$$

where the subscripts are as defined above. Note that this gives R^2.

The application of these formulas sometimes leads to surprising results. A two-choice numerical-operations test, a fairly simple and unique measure of the factor known as facility with numbers, should have had a scoring formula of $R - 3W$ to yield maximal validity for the selection of navigators in the AAF. Another, five-choice, numerical-operations test, should have had a weight of −2 for wrong answers. Thus the importance of accuracy was much greater than the a priori weights would have provided for. For the selection of bombardier students, the weight for wrong responses should have been about −.5 for the two-choice items and about zero for the five-choice items, for maximal validity of the test. For the bombardier criterion, accuracy was of relatively less importance than for the navigator. For still other tests, there were results deviating from a priori weighting, for example one test involving estimations of lengths or distances on a map seemed to require a *positive* weight for wrong answers, for maximal validity for pilots, indicating that speed was of

[1] Thurstone, L. L. The reliability and validity of tests. Ann Arbor: Edwards Bros., 1931. P. 80.

great importance in this test, even at the expense of accuracy. On the whole, the experience with scoring formulas tended to show that empirical formulas give validities slightly better than a priori weighting of wrong responses, with gains of the order of .02 to .03 being typical. On the whole, optimal weighting of wrongs gives increases of the order of .03 to .06 over validities for the rights scores used alone. There are some instances when the optimal weight for W is zero.

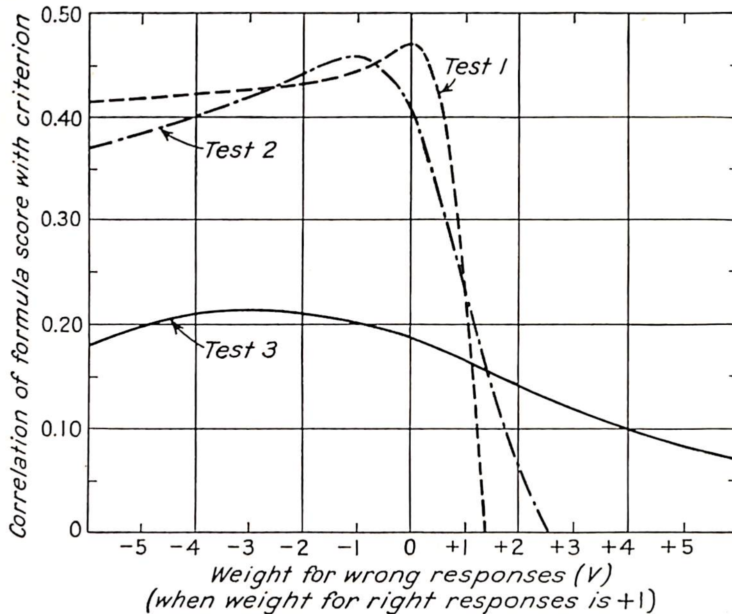


FIG. 18.5.—Practical validity of each of three tests as a function of the weight applied to wrong responses in scoring the test. Especially to be noted are the weights offering optimal validity for each test and the sensitivity of the validity to a change in weight. (*Adapted from informal AAF reports, Headquarters Training Command.*)

In Fig. 18.5 are shown the relationships between validities of formula scores in three different tests and different weights for wrongs scores in those tests when the rights scores are weighted +1. Not only can we see that there is an optimal weight for the wrongs scores for each test (.0 for test 1, −1 for test 2, and approximately −3 for test 3) but also that some weights would be detrimental to validity. These various validities can be estimated by using the correlation-of-sums formulas given in Ch. 16. The validity of each test when scored for number of right responses only can be noted at the place where $v = 0$. The amount of gain by optimal weighting can be noted by comparing this validity with the peak of the curve. There is no very marked change in validity for various negative

weights up to −.5. An error in weighting in the negative direction would apparently not be very serious. But validity drops much more rapidly if the error in the weight is in the other direction—precipitously, sometimes—if the weight goes on the positive side. Common-sense reasoning would ordinarily not permit us to choose a positive weight for the wrongs.

Empirical scoring formulas should not be derived unless samples are quite large. In some combinations of correlations among *C*, *W*, and *R*, the weight is very sensitive to minor errors in any one of the three correlations involved and may be unreasonable on the face of it. When in doubt, it is best to be conservative. It may help to plot a curve for a test, after assuming different weights for *W* and solving the correlation $R_{c.rw}$ by the formula for correlation of sums (16.25).

Factorial Validity of Right and Wrong Scores.—The procedure for maximizing practical validity for a test by using the proper scoring weights can also be applied to maximizing the correlation of a test with a factor; in other words, in increasing its loading in a factor. Recent experiences show that error scores might well be given much attention as sources of certain kinds of variance that it is worth our while to measure. Some AAF findings indicated that a trait of carefulness was quite measurable by using wrongs scores in one of several tests, whereas the number of right responses usually failed to measure it.[1] Fruchter has more recently found by factor analyzing rights scores and wrongs scores in the same tests that while the two scores in the same test may measure the same factors (in reverse), they do so to different degrees. He also found that some factors are more measurable by wrongs score than others.[2] In fact, it is possible that a certain kind of reasoning should be measured by errors rather than by correct solutions. These results have not been verified as yet, but they are suggestive of the rich possibilities there may be in fuller use and weighting of wrong responses.

Validity of Items and of Their Composite.—There have been proposals that each test be regarded as a battery and that its items be weighted according to the multiple-regression equation. The method is, of course, impractical in tests of any useful length. The result would also run counter to the goal of maximal reliability and uniqueness for each test.

There are many tests of interest and of temperament, however, in which differential weighting of items and of responses to items is common practice. This is because some items are very much more diagnostic

[1] Guilford, J. P. Printed classification tests. Ch. 25.

[2] Fruchter, B. Factorial content of right-response and wrong-response scores in a battery of experimental tests. Doctoral dissertation on file in the University of Southern California library, 1948.

of the criterion than others when they are taken alone. It is desired to give the better items full representative voice in the multiple prediction. A number of weighting procedures have been used, all of which involve some index of validity of the element (item or response). They make this much of an approach to applying the multiple-regression principles.

If we disregard the intercorrelations among items, or assume that they are approximately equal, we may apply a scoring weight that was introduced by the author. It is proportional to the correlation between item and criterion and inversely proportional to the variance of the item. The formula is[1]

$$W = \frac{p_u - p_l}{pq} + 4 \qquad \text{(Weight for an item or response)} \qquad (18.17)$$

where p_u = proportion in upper criterion group responding in a specified manner.

p_l = proportion in lower criterion group responding in the same manner.

p = proportion in the combined group so responding.

$q = 1 - p$.

This formula yields weights ranging from 0 to 8, with a weight of 4 when the item-criterion correlation is zero. The formula applies particularly in the case where the two subgroups are equal in numbers. When they do not happen to be equal, to find p_u and p_l for the two groups respectively has the effect of equalizing their importance. A standard error has been provided for the special case when the item-criterion correlation equals zero (or when $W = 4$). This is the case of the null hypothesis. The formula is

$$\sigma_{w_4} = \frac{2}{\sqrt{Npq}} \qquad \text{(Standard error of a weight when the correlation is zero)} \qquad (18.18)$$

where N = number of cases in the two subgroups combined.

p and q are as defined in the preceding equation.

Let us apply the last two formulas to a particular item. It is the question, "Would you rate yourself as an impulsive individual?" from a personality inventory that attempted to score individuals for degree of depression. From provisional scoring, the highest and lowest quarters of a group of 1,000 students had been segregated, the former being designated as the "depressed" subgroup and the latter as the "not-depressed"

[1] Guilford, J. P. A simple scoring weight for test items and its reliability. *Psychom.*, 1941, **6**, 367–374.

subgroup. Table 18.4 shows the proportions of the two subgroups responding by saying "Yes," "?" and "No" to the question. The work of solving for the weight to be assigned to each response is briefly outlined in the lower rows of Table 18.4. The three weights are 3.4, 4.3, and 4.4, for the three responses, in the order given. If we use only integral weights we should have to round them to 3, 4, and 4, respectively. Only the response "Yes" deviates far enough from 4 to be rounded to anything but 4.

TABLE 18.4.—THE SOLUTION OF SCORING WEIGHTS FOR RESPONSES TO AN INVENTORY QUESTION
($N = 500$)

	Responses			
	Yes	?	No	
p_u	.284	.180	.536	Depressed subgroup
p_l	.424	.140	.436	Not-depressed subgroup
p	.354	.160	.486	Both combined
$p_u - p_l$	−.140	+.040	+.100	
pq	.2287	.1344	.2498	
$(p_u - p_l)/pq$	−.6	+.3	+.4	
W	3.4	4.3	4.4	
Npq	114.35	67.2	124.9	
$\sqrt{Npq}$	10.69	8.20	11.18	
σ_{W_4}	.19	.24	.18	

The standard errors for deviations from 4 (null hypothesis) for the various responses are given in the bottom row of Table 18.4. Two deviations are statistically significant, but in view of the fact that only one of the responses had a weight deviating as much as barely more than half a unit from 4, it would probably be of little value to keep this item to help predict depression. It may be diagnostic of some other trait that the same weighting formula would reveal.

Although the weights proposed are all positive, which is a distinct convenience, they may be larger than is necessary in some situations. For

example, if the weights came out with a range from 2 to 6, one could deduct 2 from every weight, reducing the range from 0 to 4. This would make no difference in the effectiveness of the weights. It is their differences that are important, not their absolute amounts. The range suggested from 0 to 8 is broad enough to take care of items or responses with the highest

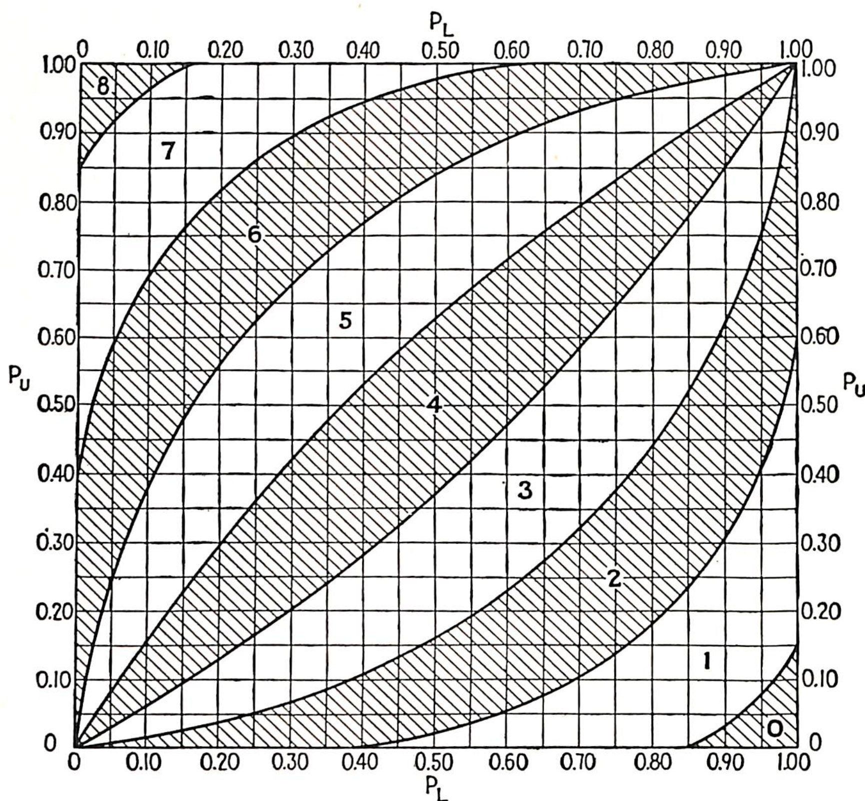


FIG. 18.6.—An abac for the graphic determinations of scoring weights for responses to test items. If 65 per cent of an upper criterion group respond in a certain manner to an item and 25 per cent of a lower group respond in the same manner, the intersection of the two lines (0.65 on the vertical scale and 0.25 on the horizontal one) falls within the band where $W = 6$. The scoring weight for this response to the item is 6.

correlations. In practice, a collection of items rarely utilizes that entire range.

For the reader's convenience, an abac for estimating integral weights on the basis of formula (18.17) is presented in Fig. 18.6. Its chief usefulness is to be found in connection with weighting responses to temperament and interest test items, but it also applies to items intended to assess abilities.

The Importance of Weighting Item Responses.—Personality tests in the past have generally leaned heavily upon a weighting system of some

kind. Typical of this statement are the Strong Vocational Interest Blank and the Bernreuter Personality Test. There are instances in which weighted scoring has materially improved reliability over that attainable with unweighted scoring. By "unweighted scoring" we mean here that responses are given a value of 0 or 1 only. Studies of validity have generally not shown much benefit from differential weighting of items. Any benefits from weighting are likely to be secured in short tests (20 items or less) only. Every test constructor, in these days of machine scoring, in which differential weighting is bothersome, should be challenged to show good cause for other than the simplest system of weighting. It is more important to be sure that items are significantly correlated with criteria and to reject those not significantly correlated. And in order to make an adequate examination of significance, we are called upon to use large samples that will serve us with reliable tests of item significance. In this way lie the selection of dependable items and the construction of dependable tests that can be safely and meaningfully applied to individuals beyond our validating sample.

Selection of Items by Correlating with an Outside Criterion.—Some tests, *e.g.*, the Strong Vocational Interest Blank, have been developed by correlating each item with an outside criterion. The outside criterion may be success in adjustment, vocational, marital, or personal. Any of the correlation methods appropriate with items may be used. Weights for scoring may be attached to responses by one of the accepted methods. The result is likely to be a valid score for the particular purpose and within the particular population on which the item validation was performed. Use of the score for other purposes and with other populations has to be defended by new empirical evidence of validity. It is probably important, also, to keep accumulating evidence of validity within the area of the test's original development.

This procedure is describable as a kind of "shotgun" approach. It gets practical results without much knowledge of why there is validity. For example, the AAF developed a Biographical Data Blank composed of items of information about the student's previous life and experiences.[1] By correlating every response to a large number of experimental items with the graduation-elimination dichotomy in pilot training and also in navigator training, two scoring keys were derived, each valid for its own purpose. One could be content with these new, unique contributions to prediction of training success. Correlational studies and factor analyses revealed that the pilot score was valid chiefly because it indicated the effectiveness of the student's background of experience in mechanical matters

[1] Guilford, J. P. (Ed.) Printed classification tests. Ch. 27.

and because it revealed his interest or motivation for pilot training. To a much smaller extent it revealed the student's status in perceptual speed and in psychomotor coordination. These were represented in the pilot criterion also. The navigator score was valid, however, primarily because it revealed the student's background experience in mathematics and to a small extent his number facility. Once the major sources of validity for each score are recognized, one is in a position to improve measurement of them. As a matter of fact, as in the example of the biographical-data approach, there often prove to be better measures of the significant factors, or better measures can be developed to replace the preliminary ones.

It is to be recognized that in an unknown sphere of prediction much progress can be made by the "shotgun" approach, of correlating a large number of items with an outside or practical criterion. It is recommended, however, that we attempt to get past this stage as soon as possible, finding out the underlying reasons for successful prediction, and improving the measuring instruments needed. Where requirements are known in terms of factorial information, the development of univocal tests is called for, and this means item-test correlation rather than item-criterion correlation.

Exercises

1. Test X has a reliability coefficient of .92, and criterion Y has a reliability of .65. Assume that the validity coefficient in four uses had values of .35, .48, .61, and .72.

   a. Determine the probable correlation between "true" measurements in test and criterion in each group.

   b. Determine the genuine validity of the fallible test in each case (assuming a perfect criterion).

2. In the preceding problem, assume that $\sigma_y = 15.0$. Compute $\sigma_{\infty x}$ and $E_{\infty x}$ for the four instances.

3. Four homogeneous tests have reliability coefficients and validity coefficients as follows:

Test	X_1	X_2	X_3	X_4
r_{xx}	.80	.80	.60	.80
r_{yx}	.70	.50	.50	.30

   a. Determine the validity coefficient in each case, assuming that each test is doubled in length.

   b. Do the same, assuming that each test is made five times as long.

   c. Do the same assuming that each test is made half as long.

4. How long (in ratio to original lengths) would it be necessary to make the four tests in Exercise 3 in order to make the validity coefficient for each one .60?

5. Assume the following data for a certain test:

$$\sigma_R = 10.0 \qquad \sigma_W = 4.0 \qquad r_{cr} = .3 \qquad r_{cw} = -.2 \qquad r_{rw} = -.4$$

a. Compute the optimal scoring weight for the wrong responses (W) when the right responses (R) are weighted +1.

b. Compute the correlation of scores obtained by the use of these weights with the criterion (C).

c. Assume in turns weights of −2.0 and +1.0 for the wrong responses and estimate the correlation with C.

6. Determine scoring weights for the three responses to the two items in Data 17B. Estimate standard errors of those weights assuming correlations of zero.

CHAPTER 19

SCALING PROCEDURES

In this chapter will be treated a number of problems of practical importance in psychological and educational measurement. The presentation must necessarily be sketchy and theory must be slighted in order to keep the material within the space of a single chapter. There is not complete agreement, as yet, concerning the validity of the procedures and as to whether they yield genuine measurements.[1] There is little doubt, however, of their practical utility and of the enlarged opportunities they offer for the treatment of data in ways that lead to profitable views and conclusions. Many forms of evaluation of objects called measurements are often very illuminating even though they do not always satisfy the logical requirements laid down by those who like precise definitions of the term *measurement*. If we use the operational term *scaling* and define the steps by which scale values are achieved, we can avoid the logical issues if we so choose. It is important to solve logical issues for the sake of clear understanding and clear thinking. We cannot postpone further use of obviously fruitful procedures, however, until all the issues are solved.

A Scale Defined.—Before going further, we should agree upon the meaning of *scale*. A scale is a linear continuum along which objects or phenomena can be located according to some abstracted property or quality. By saying "linear" we mean there are only two directions in which one can move and they are opposites: toward one pole or the other. By *continuum* is meant that there are an infinite number of positions that could be occupied on the scale. There are no gaps, and there are no natural units. The size of unit is arbitrary and is superimposed by the investigator. By *abstracted property* is meant that one aspect in which the objects of a class differ—length, weight, redness, softness, ability to perceive visual forms quickly and accurately, skill in typing, or difficulty—is singled out for observation, ignoring all others in which there may also be differences but in which we are not interested at the moment. This is

[1] See, particularly, Smith, B. O. Logical aspects of educational measurement. New York: Columbia University Press, 1938; also Gulliksen, H. Paired comparisons and the logic of measurement. *Psychol. Rev.*, 1946, **53,** 199–213.

a logical concept of a scale. There is also an operational conception of a scale. This refers to some particular description of a continuum by a series of numbers. From this point on, the term *scale* will have operational meaning unless otherwise indicated. When we need reference to the ideal scale, the word *continuum* will be employed.

Measurements Are Numerical Descriptions.—Measurements are numerical descriptions of positions on a scale. The least that numbers should mean are rank orders for the objects scaled. The order of the numbers should also be the order of the objects. There are many useful things that we can do merely with indications of rank order. There is much more that can be done when the numbers also describe distances between objects. This makes possible certain types of addition and subtraction that are meaningful, and some other operations based upon them. It is in the effort to achieve this more complete and useful type of measurement that the various scaling devices have been developed. We frequently treat scaled values as if they belong to an interval scale; as if the units are equal. The scaling process follows rational steps which to some extent justify this generosity. We can at least say that the units are rationally equal whether they may prove upon examination by some superior method to be equal or not. An example of this was seen in Ch. 12 in the scaling of test scores. If we may assume that the population is normally distributed on some continuum, then when raw scores are converted into another set of values which will yield normal distributions in large, representative samples, those scaled values should describe a numerical situation in which equal numerical distances stand for really equal differences in the quality scaled. Many of the other scaling procedures, some of which will be mentioned here, also depend upon the assumption of normal distribution in one way or another.

Some Common Scaling Methods.—In general, in psychology and education, it is reactions and judgments that are scaled. If one wishes to call judgments reactions, then the things scaled reduce to one category—reactions. Examples of reactions commonly scaled are responses to test items and their equivalents in experiments on learning, memory, apprehension, and the like. Usually, some standard of performance is set up and the datum is in the form of proportion of the time the standard is met. Proportions are sometimes used as scores (scale values) but in dealing with a single response of a certain kind, a proportion should not ordinarily be used as representing a position on a linear scale. The reasons for this will be clearer in the course of the section on scaling of test items.

Judgments that are scaled may arise in different connections; psychophysical experiments, esthetic observations, or evaluation of individuals

or of their products by inspection. Judgments fall into two general classes, comparative and absolute. These terms are only relative; probably no judgment is completely comparative or completely absolute. Comparative judgments are of the type "This sound is louder than that one," "This color is more pleasing than that one," or "This English composition shows more creative thinking than that one." An absolute judgment is of the type "I like that necktie," "This drawing rates 9 points on a scale of 10," or "That secretary is worth fifty dollars a week to her organization."

When we have the judgment "This handwriting is more legible than that," and nothing more, all we have is a statement about rank order. Both samples could be very legible, both could be very illegible, or one could be legible and the other illegible. There is not enough information in the single statement to enable us to assign each object to a scale position. If we have a number of judgments of the same type, the additional information may, at least, help us a great deal in deciding how far apart the two objects are. If only 51 per cent agreed to this judgment and 49 per cent disagreed, we would conclude that the two samples must not be very far apart. If 95 per cent agreed and 5 per cent disagreed, we could conclude that the two are widely separated. It is from such information as this that some scaling may be performed. The comparative judgment is basic to the *method of paired comparisons* and its variations.

Judgments of a series of objects in the form of linear rank order (*method of rank order*, or *order-of-merit method*) can be scaled in various ways if we permit ourselves to make certain assumptions. Another class of judgment methods can be placed in the general category of *method of successive intervals*. In any member of this group of methods, the observer has a limited number of categories in rank order. He evaluates each object by placing it within one of the categories. There is no presumption of equality of distances between neighboring categories. Coming under this heading are various *rating-scale methods*, *method of absolute judgment* or *method of single stimuli*, and even the so-called *method of equal-appearing intervals*. The latter was developed with the hypothesis that an observer can keep his categories equidistant. If we do not wish to credit an observer with this ability, we can still instruct him to strive to keep his units at equal intervals, but treat the results as if he had not succeeded or apply tests to see whether he has succeeded in following this requirement. In the following pages we will find brief accounts of how to deal with paired-comparison judgments, judgments of rank order, judgments in successive intervals, and how to scale test items. We begin with the latter.

Scaling Test Items for Difficulty

Why Scale Test Items?—A knowledge of difficulty of items serves several purposes. One of long-standing recognition is the practice of arranging items in rank order in a test, easiest items first. If there is felt a need of graduating the difficulty more or less steeply from beginning to end, with a liberal supply of items from which to draw, this need can be fulfilled. It is also important to temper the general average difficulty of items to the ability level of the group to be examined. Here several accepted rules apply, rules that are wise on both theoretical and empirical grounds. Items passed by everybody or failed by everybody are of no value for measurement purposes. This rule may be violated for the sake of introducing one or two very easy "shock absorbers" at the beginning of a test.

A rough idea of the difficulty of an item is gained from the percentage of individuals who pass or fail it. The greater the percentage of failures, the more difficult the item; the greater the percentage of passes, the easier the item. For the purpose of placing items in rank order of difficulty, the proportion passing, or failing, will do. When we want to select items at approximately equal intervals of difficulty along the scale, however, proportions will no longer suffice, as the rationale to follow will show. There are also research problems in which it is necessary to place items on a linear scale for the purpose of relating difficulty as a variable to other variables. The study of functional relationships between difficulty and other measurable features or conditions depends upon item scaling.

Rational Basis for Item Scaling.—The basic assumption for item scaling is that the proportion of the individuals passing items decreases according to the (descending) normal cumulative (integral) curve. A diagram of this is shown in Fig. 19.1. If we used proportions of *failing* examinees at each level of difficulty, the curve would be an ascending ogive. We are more accustomed to noting proportions of successes, however, so the descending curve will represent relationship accordingly. In Fig. 19.1, two items, one very easy and one of more than average difficulty, are shown. The one is passed by .92 of the examinees who attempted it and the other by .24. Knowing these two facts alone, one might be tempted to say that the first item is almost four times as easy as the second because almost four times as many passed it. This would be an erroneous statement for at least two reasons. One is that to speak of a ratio at all, one must have an absolute zero. Whether we are talking about zero per cent passing or zero per cent failing, there is no knowing that the per cent would not be something above zero if we extended the size of the sample.

The other fault with the statement is that proportions are treated as if they were additive values on a scale of equal units. Like centiles, proportions represent *areas* of population (remember they are derived from frequencies) and not positions on a linear scale. The basic hypothesis for the purpose of item scaling is that the proportions represent areas

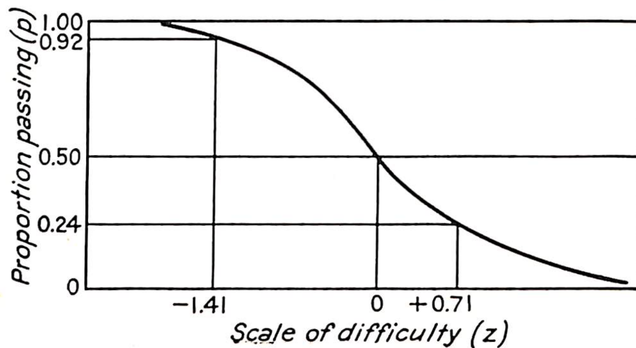


Fig. 19.1.—Proportion answering a test item correctly as a function of its difficulty. Assuming a descending ogive function, for any given proportion, there can be found, from the normal-curve tables, a corresponding value in σ units.

under segments of the normal distribution curve. Corresponding to an area of .92 in the unit normal curve is a z value of -1.4. The sign is negative because the .92 portion is the area of passing individuals and this area is above the point of difficulty of the item. For the item with .24 passing it, the corresponding z is $+.71$; positive because with less than

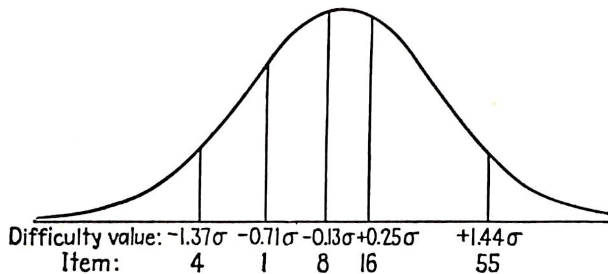


Fig. 19.2.—Another illustration of the relation of scale values of items to areas under the normal curve. Five items from an algebra examination are represented (see Table 19.1).

half passing the item they are above a point of difficulty higher than the median. Items of median difficulty are passed and failed by proportions of .50.

Figure 19.2 shows the same principles with the noncumulative normal curve. Five items from an algebra examination are represented. The area under the curve to the right of the vertical line erected at the difficulty level for each item represents the proportion passing the item.

Relation of Item Values to the Population.—Any item-difficulty scale is relative to the general picture of ability of the group—their median ability and their dispersion of ability. On the basis of the group of individuals represented in Fig. 19.2, a group of freshman engineers, item 1 had a z score of -0.71; item 4, a value of -1.37; and item 55, a value of $+1.44$. Any item with a negative sign is easier than the item of median difficulty for the group, *i.e.*, is passed by more than 50 per cent, and any item with a positive sign is more difficult than the item of median difficulty and is passed by less than 50 per cent.

TABLE 19.1.—DETERMINATION OF ITEM DIFFICULTY OF FIVE ITEMS IN AN ALGEBRA EXAMINATION FROM THE PROPORTIONS OF SUCCESSES IN THREE GROUPS OF STUDENTS

Item number	Relatively homogeneous group of high ability		Relatively homogeneous group of low ability		Large heterogeneous group of students	
	Proportion passing	z	Proportion passing	z	Proportion passing	z
1	.95	−1.64	.57	−0.18	.760	−0.71
4	.98	−2.05	.85	−1.04	.915	−1.37
8	.86	−1.08	.24	+0.71	.550	−0.13
16	.75	−0.67	.05	+1.64	.400	+0.25
55	.15	+1.04	.00		.075	+1.44

The steps in scaling items are outlined in Table 19.1. They are as follows:

Step 1. Determine the proportion of the group passing the item.
Step 2. Look up the corresponding z score in the tables. If the proportion was greater than .50, give z a negative sign; if less than .50, give it a positive sign.

In Table 19.1, the same five items are scaled on the basis of the responses of three different groups of students. The first group is made up of students of high ability in algebra, the second group is made up of low-ranking students, and the third group includes students of widely varying ability. As should be expected, the scale values of the items are generally lowest for the high-scoring group and highest for the low-scoring group. This illustrates the fact that the item values are numerically dependent upon the level of ability of the tested group used as our basis of scaling. It will be noticed, also, that the total *range* of scale values is least for the total group (which has the widest dispersion of ability), and the range is

greatest for the two groups of low variability. This is not a defect of the scaling procedure but shows the importance of not comparing difficulty values unless they come from the same population, or very similar populations, or unless adjustments are made to equate units and zero points for the different groups.

The process of converting one set of values for the same items into a scale on which they would have the same mean and standard deviation as values obtained from a different sample is like that proposed for ratings (see Table 19.15). Incidentally, it will be seen that when the proportion is zero (also when it is 1.0), a scale value cannot be estimated, as was true for the low-scoring students with item 55, which no one in this group passed. It is usually undesirable to scale items when p exceeds .97 or is less than .03, owing to the uncertainty of z in these regions. A very slight error in p is reflected by a relatively large error in z.

Correction for Chance Success.—The items just scaled were from an algebra test in which the student had to find the answers for himself. Unlike a true-false or a multiple-choice test where the student chooses one out of two or more alternatives, there is almost no chance of success by guessing. In the multiple-choice type of test, however, lucky guessing increases a student's score considerably and also increases the proportion of passing individuals. Proportions that are thus artificially inflated because of the factor of guessing should not ordinarily be used for scaling purposes unless corrections for guessing are made. The correction process is made by means of the formula[1]

$$_cp = \frac{np - 1}{n - 1} \quad \text{(Proportion of passing individuals corrected for chance success)} \qquad (19.1)$$

where n = number of alternative responses for each item.

p = obtained, uncorrected proportion of passes.

This formula is based upon the same reasoning as a priori scoring formulas which were discussed in the preceding chapter. The same assumptions and limitations apply here.

In Table 19.2 are given some illustrations of items taken from an English examination. In the first three items, the number of alternative responses was 4; in the next three, n is 3 alternatives; and in the last three, n is 2. The amount of change in correcting p is seen by comparing column (3) with column (1). In one instance (the third item), the correction yielded a proportion less than zero. Since zero success is the

[1] Guilford, J. P. The determination of item difficulty when chance success is a factor. *Psychom.*, 1936, **1**, 259–264.

TABLE 19.2.—DETERMINATION OF ITEM DIFFICULTY OF NINE ITEMS IN AN ENGLISH EXAMINATION WHEN CHANCE SUCCESS IS A FACTOR

Proportion passing	Number of alternative responses	Corrected proportion of passes	z
.90	4	.867	−1.11
.51	4	.347	+0.39
.24	4	.000	—
.42	3	.130	+1.13
.78	3	.670	−0.44
.97	3	.955	−1.70
.98	2	.960	−1.75
.79	2	.580	−0.20
.54	2	.080	+1.41

lowest actual p, some explanation is called for. It could possibly happen that the item was keyed wrong in scoring. Assuming that it was keyed properly, it was a very difficult item on which most examinees guessed the answer, and perhaps one or more of the distracters was too appealing or perhaps there was poor teaching, leaving an erroneous impression. At any rate, it is best to call such a corrected proportion zero, in which case it cannot be scaled.

An aid in correcting proportions of successes is given in the form of Table 19.3. It is recommended that where numbers of alternative responses are less than six, correction be utilized, even when scaling does not follow. The mean of uncorrected, as well as corrected, proportions gives an erroneous impression of general level of difficulty of a collection of items. For other purposes, however, such as item intercorrelations, uncorrected proportions should be used, for those are the operational values upon which correlations are based. The error variance involved will make itself felt, as it should, in attenuating correlations.[1]

MEASUREMENTS FROM JUDGMENTS OF RANK ORDER

Judgments in terms of rank order are not interval-scale measurements but if we can make certain assumptions about the actual distribution of cases along a genuine metric scale, we can transform rank orders into interval measurements. When the things ranked are individuals or the psychological or educational products of individuals and certainly when they are random samples from a normally distributed population,

[1] For more details on item scaling, see Guilford, J. P. Psychometric methods. New York: McGraw-Hill, 1936. Ch. XIII; also Davis, F. B. Item-analysis data. Cambridge: Harvard University Press, 1946.

TABLE 19.3.—A TABLE TO FACILITATE THE CORRECTION OF THE PROPORTION OF PASSING INDIVIDUALS FOR A TEST ITEM

p Uncorrected proportion	*n* Number of alternatives 2	3	4	5	*p* Uncorrected proportion	*n* Number of alternatives 2	3	4	5
.99	.980	.985	.987	.9875	.59	.180	.385	.453	.4875
.98	.960	.970	.973	.9750	.58	.160	.370	.440	.4750
.97	.940	.955	.960	.9625	.57	.140	.355	.427	.4625
.96	.920	.940	.947	.9500	.56	.120	.340	.413	.4500
.95	.900	.925	.933	.9375	.55	.100	.325	.400	.4375
.94	.880	.910	.920	.9250	.54	.080	.310	.387	.4250
.93	.860	.895	.907	.9125	.53	.060	.295	.373	.4125
.92	.840	.880	.893	.9000	.52	.040	.280	.360	.4000
.91	.820	.865	.880	.8875	.51	.020	.265	.347	.3875
.90	.800	.850	.867	.8750	.50	.000	.250	.333	.3750
.89	.780	.835	.853	.8625	.49	.000	.235	.320	.3625
.88	.760	.820	.840	.8500	.48	.000	.220	.307	.3500
.87	.740	.805	.827	.8375	.47	....	.205	.293	.3375
.86	.720	.790	.813	.8250	.46	....	.190	.280	.3250
.85	.700	.775	.800	.8125	.45	....	.175	.267	.3125
.84	.680	.760	.787	.8000	.44	....	.160	.253	.3000
.83	.660	.745	.773	.7875	.43	....	.145	.240	.2875
.82	.640	.730	.760	.7750	.42	....	.130	.227	.2750
.81	.620	.715	.747	.7625	.41	....	.115	.213	.2625
.80	.600	.700	.733	.7500	.40	....	.100	.200	.2500
.79	.580	.685	.720	.7375	.39	....	.085	.187	.2375
.78	.560	.670	.707	.7250	.38	....	.070	.173	.2250
.77	.540	.655	.693	.7125	.37	....	.055	.160	.2125
.76	.520	.640	.680	.7000	.36	....	.040	.147	.2000
.75	.500	.625	.667	.6875	.35	....	.025	.133	.1875
.74	.480	.610	.653	.6750	.34	....	.010	.120	.1750
.73	.460	.595	.640	.6625	.33	....	.000	.107	.1625
.72	.440	.580	.627	.6500	.32	....	.000	.093	.1500
.71	.420	.565	.613	.6375	.31	....	.000	.080	.1375
.70	.400	.550	.600	.6250	.30	....	....	.067	.1250
.69	.380	.535	.587	.6125	.29	....	....	.053	.1125
.68	.360	.520	.573	.6000	.28	....	....	.040	.1000
.67	.340	.505	.560	.5875	.27	....	....	.027	.0875
.66	.320	.490	.547	.5750	.26	....	....	.013	.0750
.65	.300	.475	.533	.5625	.25	....	....	.000	.0625
.64	.280	.460	.520	.5500	.24	....	....	.000	.0500
.63	.260	.445	.507	.5375	.23	....	....	.000	.0375
.62	.240	.430	.493	.5250	.22	....	....	....	.0250
.61	.220	.415	.480	.5125	.21	....	....	....	.0125
.60	.200	.400	.467	.5000	.20	....	....	....	.0000

we can assume that the sample approximates a normal distribution. At any rate, it is often safe to assume that individuals or samples near the center of the range are less far apart in quality than those near the extremities. Because this is one of the important characteristics of the normal distribution, except for possible instances of very marked skewing in the sample distribution, the assumption of normality probably does little violence to the data.

Transforming Ranks into Centile Positions.—From a rank position assigned to any specimen by a judge, we can fortunately specify its centile position, and from its centile position we can determine its equivalent standard-score rating or any other rating based upon standard scores. This is the underlying principle of the most commonly used procedure of deriving measurements from ranks.

Usually a rank of 1 means the top individual or specimen in the group, and where there are n things ranked, the lowest item receives a rank of n. The formula for computing the centile position corresponding to a rank is

$$P = 100 \times \frac{n - r + 0.5}{n} \quad \text{(A centile position derived from a rank)} \quad (19.2)$$

where n = number of things ranked.
r = particular rank assigned to one thing.

In a set of 10 things ranked, for example, a rank of 1 would have the centile position of 95.0, as derived from the formula

$$P = 100 \times \frac{10 - 1 + 0.5}{10} = 100 \times \frac{9.5}{10} = 100 \times 0.95 = 95$$

This is reasonable, when we remember that the top person is conceived as occupying a range that includes the highest tenth of the scale of 100 centile ranks, or from 90 to 100. The midpoint of this range is 95. The top person in a group of 15 would have a centile position of 96.7, for

$$P = 100 \times \frac{15 - 1 + 0.5}{15} = 100 \times \frac{14.5}{15} = 96.7$$

Similarly, a rank of 2 in 15 has a centile position of 90.0, and a rank of 7 in 15, a position of 56.7. Table 19.5 gives the solution of all ranks from 1 to 15 in a set of 15 things ranked. Column (2) presents the numerators of the fraction in formula (19.2) and column (3), the centile positions.

We could now determine the standard scores corresponding to these centile positions, but having previously found fault with the practical use of standard scores (see Ch. 12) we shall here recommend the use either of T scores or C scores. The corresponding T scores are looked up in

TABLE 19.4.—CONVERSION TABLE TO FACILITATE THE TRANSLATION OF RANK ORDERS INTO C-SCALE MEASUREMENTS

Centile-position Range for Each C-Score Unit	*C Score*
98.9+	10
96.1–98.8	9
89.5–96.0	8
77.4–89.4	7
60.0–77.3	6
40.2–59.9	5
22.8–40.1	4
10.7–22.7	3
4.1–10.6	2
1.3– 4.0	1
0– 1.2	0

Table 12.3, and they are listed in column (4) of Table 19.5. The *C* scores corresponding to the centile positions are conveniently looked up in Table 19.4, especially provided for this purpose. In the opinion of the author, most rank orders from individual judges are so subject to errors of observation that any scale finer than the *C* scale for expressing them should be out of the question.

Scaling Ranked Data When Distributions Are Not Normal.—When it is strongly suspected or known that the distribution of cases is not normal, the procedures just described should be seriously modified or replaced. Other procedures not assuming anything about the form of distribution of the things ranked have been described elsewhere by the author.[1]

Whenever the form of distribution is fairly well known, certain other procedures are in order. For example, the method of rank order is well adapted to the evaluation of English compositions or themes. Although any single teacher's usual grading of themes is notoriously subjective and faulty, he can place them in rank order for excellence, as judged for certain adopted criteria, which would probably correlate well with another judge's rank order. If it is known that the distribution of scholastic aptitude, or better yet, the distribution of English achievement, is positively skewed, the grading of the themes can then be planned accordingly. The known distribution will tell the teacher what number of A's, B's, C's, etc., should be expected, and having the papers in rank order, he can proceed to assign the expected number of A's to those heading the list, next the expected number of B's, etc. The distribution of marks for the themes

[1] Guilford, J. P., *op. cit.*, Ch. VIII.

TABLE 19.5.—DETERMINING T SCORES AND C SCORES FROM RANK ORDERS

(1) Rank	(2) Number the rank exceeds or is equal to	(3) Centile position or cumulative percentage to the mid-rank	(4) *T* score (from Table 12.3)	(5) *C* score (from Table 19.4)
1	14.5	96.7	68	9
2	13.5	90.0	63	8
3	12.5	83.3	60	7
4	11.5	76.7	57	6
5	10.5	70.0	55	6
6	9.5	63.3	53	6
7	8.5	56.7	52	5
8	7.5	50.0	50	5
9	6.5	43.3	48	5
10	5.5	36.7	47	4
11	4.5	30.0	45	4
12	3.5	23.3	43	4
13	2.5	16.7	40	3
14	1.5	10.0	37	2
15	0.5	3.3	32	1

will then coincide in mean, variability, and form with the known distribution of the class. This is far better procedure than to adopt the outworn procedure of assuming that every class has a normal distribution and to assign the same percentage of A's, B's, C's, and even F's to every class, regardless of its selection, the kind of teaching received, or the actual achievement.

SCALING JUDGMENTS FROM PAIRED COMPARISONS

The Fullerton-Cattell Principle.—As a consequence of extensive experiments in psychophysical problems, Fullerton and Cattell came to a conclusion that has been basic to much scaling activity, particularly in education.[1] The conclusion was that equally often noticed differences are equal, unless always or never noticed. If a lifted weight *B* is judged heavier than weight *A* 85 per cent of the time and weight *K* is judged greater than weight *J*, also 85 per cent of the time, then the psychological distance from *J* to *K* is equal to that from *A* to *B*. The equality of percentage is the basis of equating scale distances between objects. A

[1] Fullerton, G. S. and Cattell, J. McK. On the perception of small differences. *Pub. Univ. Penn., Phil. Series*, No. 2, 1892.

percentage of 75 was adopted as the unit of difference when expressed in these terms and when only one of two judgments was permitted, *e.g.*, *A* greater than *B* and *A* less than *B*. This is because a centile of 75 is one probable error distant from the mean in a normal distribution. Any other percentage difference could be converted into corresponding probable error units. This procedure was used in the early days of educational measurement to establish scales of handwriting quality, of excellence of English themes, and the like.

Thurstone's Law of Comparative Judgment.—In more recent years, the problems of comparative judgment and how to apply it in extracting scale distances from data have been more thoroughly rationalized by Thurstone.[1] He demonstrated the fact that the Fullerton-Cattell principle holds only under certain conditions and gave a more general solution to the problem. Although in the scaling methods to be described here we will limit ourselves to the assumption of this principle, it will pay us to look into Thurstone's rationale in order to obtain a clearer grasp of the problem.

Thurstone begins by assuming that in the observation of any object in any of its aspects, repeated perceived quantities will differ. There is a characteristic scale (continuum) position for each object which could be regarded as its mean or median or modal value but there is a dispersion of other positions the observation may take. This holds true whether we are thinking of the same observer at different times or of a number of observers who note the object only once. The dispersion of observations may be assumed to be normal. This assumption is in line with the long-standing principle in the natural sciences that errors of observation are normally distributed. It is a common assumption in all psychophysical practice and theory. The phenomenon can readily be tested empirically by asking an observer to draw a straight line to reproduce one he inspects and to repeat this until a large number of independent reproductions have been made. The frequency distribution of the reproductions may be taken in large part as the observer's fluctuating perception (we will ignore for the moment his variable errors in drawing.)

This phenomenon is pictured in Fig. 19.3. The first two distributions are for the observations of some one aspect of two objects, *A* and *B*. Object *B* has a higher mean position on the continuum (which is a purely psychological continuum) than object *A*. These mean positions we will call S_b and S_a, respectively. If the two objects have equal dispersions as denoted by equal standard deviations, that is, if $\sigma_a = \sigma_b$, either σ might be used as the unit on the psychological scale. We could say that S_b is

[1] Thurstone, L. L. Psychophysical analysis. *Amer. J. Psychol.*, 1927, **38**, 368–389.

at a point equal to $+1.5\sigma_a$ on the scale of object A, or that S_a is at a point $-1.5\sigma_b$ on the scale of object B. Since it is the same continuum, both distances are on the same scale and it is arbitrary which mean we use as a reference point and which unit we use for scaling purposes, even if the two σ's were not equal.

Comparative judgments of objects A and B will not tell us immediately about their separations in terms of either σ_a or σ_b. Comparative judgements refer to *differences* between objects rather than to single objects.

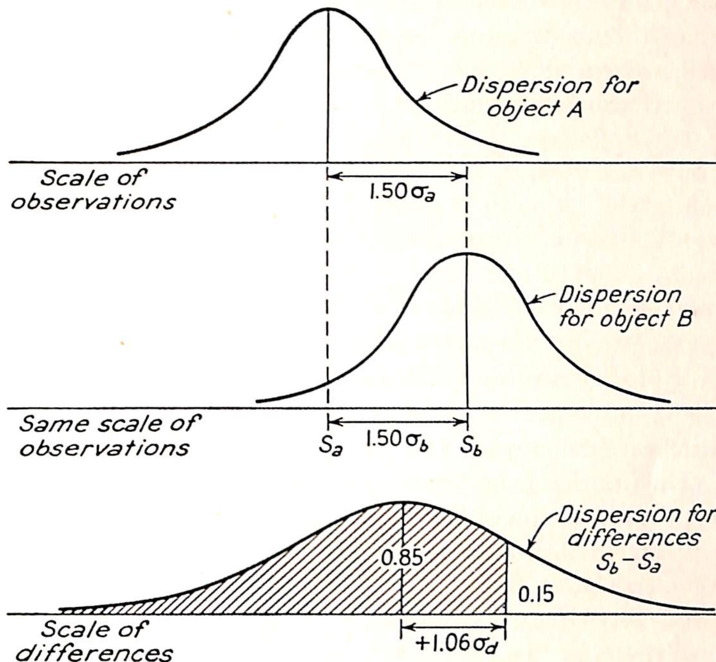


FIG. 19.3.—Mean scale positions and dispersions of two observed objects on a psychological scale, also a distribution of the observed differences of scale positions.

From Fig. 19.3 it is clear that as both A and B fluctuate in value from time to time, at any one moment their relative positions may be such that the judgment is "B decidedly greater than A," at another moment it is "B slightly less than A." It will depend upon whether A is perhaps very high on the scale at the moment and whether B is at the same time high, medium, or low, within its own range. If we take a number of repeated comparative judgments, each one representing a certain scale difference between the two objects, we would have a frequency distribution of such differences. If we restrict ourselves, as we usually do, experimentally, to two categories of judgment, all the information we have is the proportion of the time B is judged greater than A (or its complement,

the proportion of the time *A* is judged greater than *B*.) If we assume a normal distribution for these differences, we have a picture like the third in Fig. 19.3. If the distributions for *A* and *B* singly are normal, the distribution of the differences will be normal.

Let us say that we found by experiment .855 of the judgments were "*B* greater than *A*." This is represented by the shaded area under the normal curve. The standard measure, or z, corresponding to .855 is +1.06. This value is in terms of the standard deviation of the differences as the unit. If we call this distance z_{ab}, meaning that it is the standard measure corresponding to the comparative judgment in which objects *A* and *B* are involved, we are now ready to state Thurstone's law of comparative judgment in terms of an equation:

$$S_b - S_a = z_{ab}\sqrt{\sigma^2_a + \sigma^2_b - 2r_{ab}\sigma_a\sigma_b} \quad \text{(Law of comparative judgment)} \qquad (19.3)$$

where $S_b - S_a$ = scale separation of objects *A* and *B* on a selected continuum.

σ_a and σ_b = dispersions of objects *A* and *B* on the same continuum.

z_{ab} = standard-measure distance between *A* and *B*.

r_{ab} = correlation between pairs of positions of *A* and *B* at the moment of comparison.

The entire expression with the radical is recognized as the standard deviation of a difference between two measures, in other words, $\sigma_{(b-a)}$. It measures the dispersion of the differences $B - A$ in this situation. Its size is derivable from σ_a and σ_b with the covariance term that depends upon the correlation of *B* and *A*. The entire expression for σ_d is in the equation to denote that $\sigma_{(b-a)}$ is the unit of the scale measuring $S_b - S_a$ if we stop operations at that point, that is, deriving z_{ab} from the proportion of judgments, p. We could write equation (19.3) as $S_b - S_a = z_{ab}\sigma_{(b-a)}$. We could write the equation also as simply $S_b - S_a = z_{ab}$, except for the fact that there are other pairs of objects compared: *A* and *C*, *B* and *C*, *C* and *D*, and so on. For each pair we could derive a z value, but its unit of measurement would depend upon the particular σ_d of the particular pair: z_{ac} would have as its unit $\sigma_{(c-a)}$, z_{bc} would have its unit $\sigma_{(c-b)}$, and so on. If we want a constant unit for all separations between objects, either we have to be satisfied that the σ's of the differences are all equal or we have to make adjustments where necessary to achieve a common unit. The σ's of the differences would all be equal if the σ's of all the dispersions of single observed objects were all equal and if the correlations of pairs were all equal.

The correlation term in equation (19.3) has to do with the interdependence of judgments. If when observing *A* as being unusually large we also

tend to observe B as large, the correlation r_{ab} would be positive. If there are contrast effects, so that as A appears smaller B appears larger, and vice versa, r_{ab} would be negative. If the appearance of A can be anything in its distribution regardless of what B appears to be, r_{ab} is zero. The simplest assumption to make is that $r_{ab} = .00$; that observed values of objects are independent. This eliminates the covariance term from the equation and it becomes

$$S_b - S_a = z_{ab}\sqrt{\sigma^2_a + \sigma^2_b} \quad \text{(Law of comparative judgment when observations are independent)} \qquad (19.4)$$

If we make the additional assumption that $\sigma_a = \sigma_b$, we can substitute σ_a for σ_b and the equation becomes

$$S_b - S_a = z_{ab}\sqrt{2\sigma^2_a} = z_{ab}\sigma_a\sqrt{2} \quad \text{(Same, as 19.4 assuming also equal dispersions for objects)} \qquad (10.5)$$

If we extend the assumption to all objects compared, we may drop the σ_a, since it is the unit of measurement, and the workable formula becomes

$$S_b - S_a = z_{ab}\sqrt{2} \qquad (19.6)$$

It is only under the conditions assumed in formula (19.6) that the Fullerton-Cattell principle holds. If dispersions of observations of single objects are not equal or if correlations are not equal, the unit changes and the same proportion of judgments might mean different scale separations between different pairs. In the method to be described below, we will make the necessary assumptions to fit the principle and solve the scaling problem on that basis. We will then make a test of the internal consistency of the scale values of the objects to determine whether or not they support the assumptions.

The Solution of a Scaling Problem.—The data used to illustrate the scaling process are given in Table 19.6. These data were extracted, with

TABLE 19.6.—PROPORTIONS OF THE TIMES THAT THE NAME REPRESENTED AT THE TOP OF THE TABLE WAS JUDGED PREFERABLE TO THOSE AT THE SIDE

Name	Name number					
	1	2	3	4	5	6
1. Robert	(.50)	.30	.23	.08	.04	.03
2. Jack	.70	(.50)	.42	.20	.09	.07
3. George	.77	.58	(.50)	.26	.12	.11
4. Harry	.92	.80	.74	(.50)	.33	.28
5. Henry	.96	.91	.88	.67	(.50)	.45
6. Albert	.97	.93	.89	.72	.55	(.50)

some estimating, from a report made by Walton on a study of preferences for boys' first names.[1] Eighteen common boys' names had been presented for paired comparisons to 108 male university students. The first successive steps in scaling are represented in Table 19.7. Those steps are:

TABLE 19.7.—SCALE SEPARATIONS BETWEEN NAMES AT THE TOP AND NAMES AT THE SIDE, ON THE SCALE OF THE STANDARD DEVIATION OF A DIFFERENCE; ALSO DERIVED SCALE VALUES

Name	Name number					
	1	2	3	4	5	6
1. Robert	.000	− .524	− .739	−1.405	−1.751	−1.881
2. Jack	+ .524	.000	− .202	− .842	−1.341	−1.476
3. George	+ .739	+ .202	.000	− .643	−1.175	−1.226
4. Harry	+1.405	+ .842	+ .643	.000	− .440	− .583
5. Henry	+1.751	+1.341	+1.175	+ .440	.000	− .126
6. Albert	+1.881	+1.476	+1.226	+ .583	+ .126	.000
Σz	+6.300	+3.337	+2.103	−1.867	−4.581	−5.292
M_z	+1.050	+ .556	+ .350	− .311	− .764	− .882
$M_z\sqrt{2}$	+1.48	+ .79	+ .49	− .44	−1.08	−1.25
$(M_z\sqrt{2} + 1.25)$	+2.73	+2.04	+1.74	+ .81	+ .17	.00
$(M_z\sqrt{2} + .40)$	+1.45	+1.19	+ .89	− .04	− .68	− .85

Step 1. From an inspection of the proportions of judgments, decide upon the rank order of the objects. The sums of columns will be sufficient evidence of rank order. Place the object with the highest probable scale value at the left and the one with the lowest at the right. The (.50) in each diagonal cell is an assumed value.

Step 2. Convert every proportion into a corresponding z from tables of the unit normal distribution. These are listed in the body of Table 19.7. The z's in the upper-right part of the table are identical numerically with those in the lower-left part; only algebraic signs differ. It is well to find both sets from the normal-curve table, for checking purposes.

Step 3. Sum the columns of scale separations. This is preparatory to finding the mean of each column.

Step 4. Compute the mean of each column. This is a temporary scale value for each object. The unit of this temporary scale is the σ of a difference.

[1] Walton, W. E. The affective value of first names. *J. appl. Psychol.*, 1937, **21**, 396–409.

Step 5. Convert the temporary scale values into those having a new unit which is the σ of the dispersion of any one object on the continuum of observation. The ratio of this unit to the temporary one is $\sqrt{2}$. The new scale values are $M_z \sqrt{2}$.

Step 6. (First alternative) This step and the next one are concerned with the question of where the zero point should be. The set of scale values found in step (4) has reference to a zero point that has no psychological significance. It is a function of the particular set of data used. There are two changes in zero point, both arbitrary but one psychologically meaningful in this problem. The first suggestion is aimed merely at getting rid of negative values. We may arbitrarily decide to let zero be the value of the lowest object in the list. This can be accomplished by adding to all scale values an amount numerically equal to the lowest negative value; in this case the constant is 1.25. Adding this constant, the range extends from .00 to 2.73. Before it was from -1.25 to $+1.48$. The range is the same; the zero point is at a different place in the range.

Step 7. (Second alternative) In the case of affective scales, or other bipolar scales, there is a natural zero point dividing positive and negative reactions. For the affective scale it is an indifference point. This point cannot be established from paired-comparison data alone, but a supplementary procedure will give us the necessary data. This procedure will be described in the next paragraphs.

A Method of Absolute Judgments for Affective Scales.—The same judges who gave the paired-comparison reactions were asked in another set of observations to react to each name separately. They were asked regarding each name, "Do you like it?" The answers were, in effect, "Yes" or "No." The datum for each object is the proportion of the observers saying "Yes." We may let these proportions also be represented by areas under the normal curve and derive the corresponding z value. The data for the six names of our illustration are given in Table 19.8. For each name is given its p_l (proportion of likes) and from this information, its z_l. If a z is positive (when p_l exceeds .50) we may say that it is on the "pleasure" side of the indifference point; if negative, on the "unpleasure" side of the indifference point.

These z values in themselves might be used as scale positions, and they are with reference to a meaningful zero point without anything further. There would be a question as to whether the units for all values are equal,

and there is no way of testing this, as there is for paired-comparison scale values. In practice, too, if any object receives a proportion of 1.00 for likes and if any receives a proportion of .00, either would be unscalable by this method. There are many objects that would receive such unanimous votes. But we can use the results in Table 19.8 to locate a meaningful zero point for the paired-comparisons scale values.

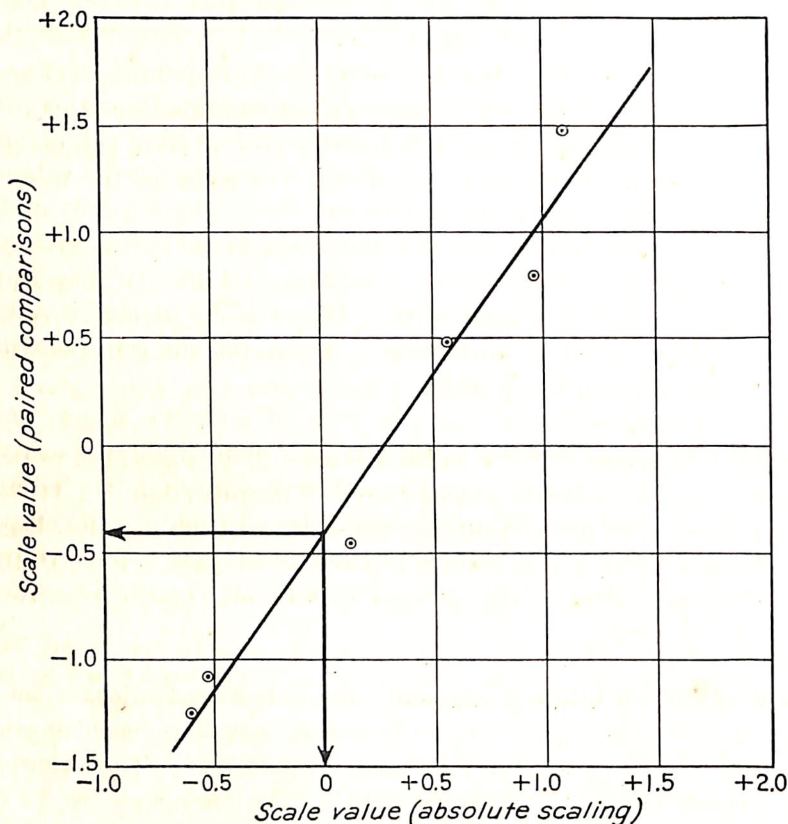


FIG. 19.4.—Linear relationship between scale values obtained by paired comparisons and those obtained by absolute scaling, showing the location of the indifference point on the former scale corresponding to the judged indifference point on the latter.

The scale values for the six names as found by the two procedures should be linearly related and highly correlated. They are, as Fig. 19.4 shows. There the paired-comparison values, noted as $M_z\sqrt{2}$ in Table 19.7, have been plotted as a regression upon the z_l values of Table 19.8. With a very close fit of observed points to the regression line, a graphic solution will suffice for finding what value in the $M_z\sqrt{2}$ scale is equivalent to a z_l equal to zero. A regression line can also be described by an equation found as explained in Ch. 15. Such an equation here turns out

TABLE 19.8.—SCALE VALUES FOR NAMES DERIVED FROM ABSOLUTE JUDGMENTS

	Name number					
	1	2	3	4	5	6
p_l	.864	.835	.717	.554	.298	.270
z_l	+1.10	+.97	+.57	+.14	−.53	−.61

to be $Y' = 1.46X - .40$. When $z_l = 0$, the corresponding value on the other scale is −.40. The correlation coefficient in this case is .98. Where it is much less than that, it would be better procedure to assume a correlation of +1.0 in setting up a transformation equation as described in the last part of this chapter (Table 19.15).

Knowing that the zero point (for indifference) should be at −.40, we can locate all six of the boys' names accordingly by adding .40 to all values given in the row with the symbol $M_z\sqrt{2}$. The range now extends from −.85 for Albert to +1.45 for Robert. Where the name Harry originally had the appearance of being disliked it now seems to be just about at the indifference point.

Internal Consistency of the Scale Values.—It is important to examine the scale values for internal consistency. This procedure is a fairly good check as to whether our simplifying assumptions were sound. The operations of the checking process are illustrated in Tables 19.9, 19.10, and 19.11. The principle of the process is to work backwards from the

TABLE 19.9.—SCALE SEPARATIONS CALLED FOR BETWEEN NAMES AT THE TOP AND THOSE AT THE SIDE BY THE SCALE VALUES ON THE DIFFERENCE SCALE

Name	Name number					
	1	2	3	4	5	6
1. Robert	.000	− .494	− .700	−1.361	−1.814	−1.932
2. Jack	+ .494	.000	− .206	− .867	−1.320	−1.438
3. George	+ .700	+ .206	.000	− .661	−1.114	−1.232
4. Harry	+1.361	+ .867	+ .661	.000	− .453	− .571
5. Henry..........	+1.814	+1.320	+1.114	+ .453	.000	− .118
6. Albert..........	+1.932	+1.438	+1.232	+ .571	+ .118	.000

obtained scale separations to the proportions of judgments that these separations would call for. We then see whether these expected proportions agree with the obtained ones. In a sense, the expected proportions represent our hypothesis, and deviations of the obtained proportions

TABLE 19.10.—PROPORTIONS CALLED FOR BY THE SCALE SEPARATIONS IN TABLE 19.9

Name number	1	2	3	4	5
2	.69				
3	.76	.58			
4	.91	.81	.75		
5	.96	91	.87	.67	
6	.97	.93	.89	.72	.55

TABLE 19.11.—DISCREPANCIES BETWEEN OBSERVED AND EXPECTED PROPORTIONS

Name number	1	2	3	4	5
2	+.01				
3	+.01	.00			
4	+.01	−.01	−.01		
5	.00	.00	+.01	.00	
6	.00	.00	.00	.00	.00

from them can be examined for significance by t tests. If we have to reject the null hypothesis anywhere in the data, we must also reject our assumptions at that point. The steps are as follows:

Step 1. Using the M_z values in Table 19.7, find the corresponding scale separations between all pairs of objects. These are found by pairing every possible pair of objects and subtracting one M_z from the other. These "expected" separations are given in Table 19.9. The unit is a hypothetical common σ of a difference.

Step 2. From the z values in Table 19.9, find from the table of the unit normal curve each corresponding expected proportion. These are given in Table 19.10. It may be well for checking purposes to do the step for the entire table, but only one set of proportions need be recorded.

Step 3. Find each discrepancy between the obtained proportion (from Table 19.6) and its corresponding expected proportion (from Table 19.10). These are listed in Table 19.11. The few discrepancies are no greater than .01. Probably all are smaller than the corresponding σ_p standard errors. We can test one that has probably the smallest standard error and see. This would be the expected proportion of .91 for the difference between Robert and Harry. With an N of 108, the $\sigma_p = .028$. The deviation is .01, which gives a t of 0.36. There is no reason to reject

the idea that the obtained proportion of .92 was a chance deviation from .91.

If the internal-consistency test had shown that the assumptions of equal dispersions of observations of objects and of independence of correlations were unjustified, we can still scale the objects by special steps which we cannot go into here. The steps are described elsewhere.[1]

Scaling Judgments in Successive Categories

Persons or things are sometimes judged by being assigned to a small number of defined classes. This is true of rating-scale methods and also of the method of equal-appearing intervals. Although the points on the rating scale and the groups in the method of equal-appearing intervals are presumably equidistant in the minds of the raters, it is often best to treat them as merely *successive* intervals on the scale. In the use of successive intervals or groups is latent the idea of rank order. But in scaling those groups, we do not treat them just as we did ranks. For one reason, there are too few "ranks" in this instance, and for another reason, too many specimens are judged virtually equal by being placed in the same category. In the ranking method, we would have had full rank-order information about these things now seemingly equal within the same category.

Likert's Scaling Procedure.—There are several procedures by which the scaling of things judged in successive intervals can be accomplished. We shall describe one procedure here that has been given prominence by Likert.[2] It assumes a normal distribution of the things rated. As an example, let us take the case in which the common word "recklessness" was judged by 400 students on a scale of five categories in the order: "very unpleasant," "unpleasant," "indifferent," "pleasant," and "very pleasant." The number of students who rated the word in each category may be seen in Table 19.12. Here it is seen that 58 reacted by marking the word "very unpleasant"; 185 marked it "unpleasant"; 104, "indifferent"; 48, "pleasant"; and 5, "very pleasant." Along the base line of our distribution curve will be placed the so-called *affective scale*, which represents a continuum extending from the most unpleasant experience at the left to the most pleasant experience at the right, with a point of absolute indifference at the center. We shall not assume that the five descriptive terms used to describe the five categories are really equidistant

[1] Guilford, J. P. Psychometric methods. Ch. VII.

[2] Likert, R. A technique for the measurement of attitudes. *Arch. Psychol.*, 1932, No. 140.

TABLE 19.12.—THE CALCULATION OF THE AVERAGE STANDARD MEASUREMENT FOR JUDGMENTS IN ONE OF SUCCESSIVE CATEGORIES

EXAMPLE: DISTRIBUTION OF JUDGMENTS OF 400 STUDENTS ON PLEASANTNESS AND UNPLEASANTNESS OF THE WORD "RECKLESSNESS"

Categories	Very unpleasant	Unpleasant	Indifferent	Pleasant	Very pleasant
Frequencies	58	185	104	48	5
Proportions ($p_2 - p_1$)	.1450	.4625	.2600	.1200	.0125
Proportions below the category (p_1)	.0000	.1450	.6075	.8675	.9875
Proportions below, plus those in the category (p_2)	.1450	.6075	.8675	.9875	1.0000
Ordinate at lower limit of category (y_1)	.0000	.2279	.3844	.2143	.0323
Ordinate at upper limit of category (y_2)	.2279	.3844	.2143	.0323	.0000
$y_1 - y_2$	− .2279	− .1565	+ .1701	+ .1820	+ .0323
z	−1.57	−0.34	+0.65	+1.52	+2.58

psychologically. We do not know what their respective spacings are. We propose to find out by the procedure next to be described.

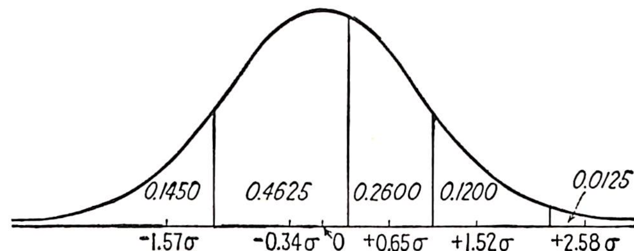


FIG. 19.5.—Segments under the normal curve representing the proportions of judgments in the five categories of pleasantness and unpleasantness. The scale value of each category is the arithmetic mean of the segment.

As usual, we let area under the normal curve stand for frequencies or proportions. Since we are going to work in terms of a curve with unit area, we must deal with proportions. The proportions of judgments in the categories (frequency divided by N in each case) are given in Table 19.12. They are illustrated by marking off five divisions under the normal curve in Fig. 19.5. It will be seen at once that the segments of the base line, our linear scale, occupied by the respective categories are not by any means equal in width. If we want a single value to stand

for each category, probably the first idea that occurs to us is to use the midpoint of each interval. There are two objections to such a choice in this situation. In the first place, the two end categories stretch off to indefinite limits, unless we arbitrarily assign definite outer limits to the end categories. In the second place, the mean of the cases within each group would best represent the cases within the interval, and this will not coincide with the midpoint. We have previously allowed the midpoint of an interval to represent the cases within that interval, but in those instances we had more than 10 intervals as a rule, and they were narrower in range. Here we have only 5. There is always a small error introduced by using the midpoint to stand for all the cases, and the coarser the grouping (the smaller the number of classes), the greater is this error. Here we feel compelled to compute the mean of each group. The mean of any segment under the normal curve between two limits is given by the formula

$$z = \frac{y_1 - y_2}{p_2 - p_1} \quad \text{(Mean of segment under the normal curve)} \quad (19.7)$$

where z = z score or standard measurement in terms of which the mean is given.

y_1 = ordinate at the *lower* limit of the segment and is to be determined from Table C.

y_2 = ordinate at the *upper* limit of the segment.

p_1 = proportion of cases *below the segment.*

p_2 = proportion of cases *below the upper limit of the segment.*

$p_2 - p_1$ = proportion *within the segment.*

Figure 19.6 illustrates these symbols.

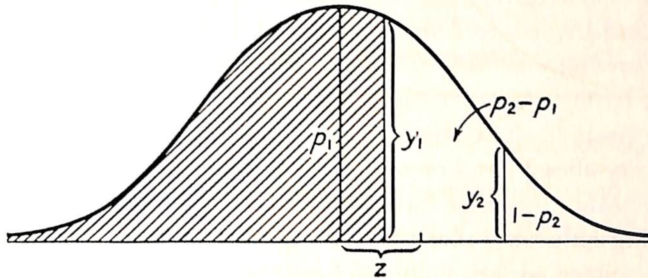


FIG. 19.6.—Illustration of the constants represented in formula (19.7) and their relation to the normal curve.

The work of computing the mean of a segment is completely illustrated in Table 19.12. First are listed the five proportions within the five segments, then the proportions p_1 and p_2. Next, from Table C, which is entered with p_1 and then p_2, are found the corresponding ordinates. The

next step is to find the differences in ordinate $y_1 - y_2$. Lastly formula (19.7) can be applied, giving the z values we wanted. The z values are listed in the last row of Table 19.12 and are shown graphically in Fig. 19.5.

It can be seen now that the scale separations between successive categories are not equal. From the point for "very unpleasant" to "unpleasant" is a standard distance of 1.23 (*i.e.*, from −1.57 to −0.34). From "unpleasant" to "indifferent" is a distance of 0.99 standard unit. From "indifferent" to "pleasant" the distance is 0.87; from "pleasant" to "very pleasant" it is 1.06 units. The discrepancies could have been worse; but they are serious enough to cause us to hesitate in labeling the original categories with the numbers 1, 2, 3, 4, and 5, as if they were just one unit apart, and computing mean scale positions for words on the basis of this scale.

A Common Scale for All Specimens.—The mean of this normal distribution for the judgments of the word "recklessness" comes at the standard score of zero, and this will be true regardless of what word or other stimulus is being judged. This does not mean that all words have the same average position on the affective scale. On the common scale upon which other words are also to be evaluated, we shall want to anchor the zero point to serve for all cases, and the most reasonable place for this is the indifference point. In this distribution, the "indifference" category came at a point 0.65σ above the mean. If we now shift the zero point up to this position as our point of reference, we shall find that the mean affective position of the word "recklessness" is therefore at −0.65.

We can similarly find the mean positions of all the words so rated on the same affective scale whose zero point is the point of indifference rating. But there are sometimes obstacles in the way. It is unlikely that distributions for all words will be normal in form. Some may be bimodal, even, and many may be skewed, particularly those words near the ends of the affective scale. Another difficulty is that the real dispersions or variabilities are probably not the same for all words. On some, the judges may agree very closely, and on others, they may differ considerably. The standard-deviation units we have for the word "recklessness" would not necessarily coincide with those for other words. It is better that we try to determine once and for all the relative spacing of the five judgments as determined by several sample distributions and use these facts, assuming that this spacing remains relatively constant no matter what word is being rated.

Following this line of thought and using the word "recklessness" and its distribution as the basis for our scale, the positions of the five categories on this common scale would be 0.65 units lower than those given in the

last row of Table 19.12. Deducting 0.65 from them, we have −2.22, −0.99, 0.0, +0.87, and +1.93 as the scale positions of the five categories. Should we wish to make the unit of the scale equal 0.1σ, the five values become −22.2, −9.9, 0.0, 8.7, and 19.3, respectively. And should we wish integers, rounded, they would become −22, −10, 0, 9, and 19. Should we wish to have all positive numbers, we could add to them a constant large enough to make them all positive. If the constant 22 is added, to make the lowest value zero, we have 0, 12, 22, 31, and 41. Actually, in this case, the increments are so near to 10 (they are 12, 10, 9, and 10) that one would almost be tempted to forget the discrepancies and revert to a 0, 1, 2, 3, and 4 scale.

One would not ordinarily permit responses to just one word to determine the spacing of the categories of judgment for all words. In this particular investigation, there were some 400 words rated. It would be wise to evaluate the separations among the five categories at least 20 times, with 20 different words as a basis. Words that yield judgments in all five categories are preferable for this purpose. This would yield at least 20 estimates of the scale separations between the neighboring categories, and a mean of these 20 estimates would give a much more adequate basis for the final scale positions of the five categories. It would be important in this to be sure that the total range of category values is about the same for all words. Since words vary in their dispersions on the affective continuum, the ranges may not be the same. Either adjustments must be made for varying range, or else those words giving ranges differing noticeably from the rest should be left out of account. Space does not permit going more completely into detail as to this procedure.

There are, in addition to the few scaling methods described here, a number of others, but they would take us beyond the scope of this introductory treatment. Here we are confined to the ones most easily applied and the otherwise most practical methods.[1]

Transforming One Distribution into Terms of Another

In Ch. 12 we encountered problems of scaling test scores to common terms—a common mean and common standard deviation, with or without normalizing the distribution. A common scale of z scores, T scores, or C scores was the result. Here we face a slightly different and more general problem. The desired mean and standard deviation might be

[1] See, particularly, Saffir, M. A. A comparative study of scales constructed by three psychophysical methods. *Psychom.*, 1937, **2**, 179–199; Mosier, C. I. A modification of the method of successive intervals. *Psychom.*, 1940, **5**, 101–107; Guilford, J. P. The computation of psychological values from judgments in absolute categories. *J. exp. Psychol.*, 1938, **22**, 32–42.

almost anything, depending upon the practical situation. We want to make the means and standard deviations of two or more distributions comparable and there is a choice of which distribution will be standard. This problem has arisen previously from time to time; for example, in connection with the scaling of test items in two or more different populations and in multiple predictions in Ch. 16. The chief interest here, however, is to find what the observed values in one distribution should become in order to arrive at the same mean and standard deviation as in a parallel distribution.

One practical instance in which this kind of transformation may be useful is in deriving school marks from examination scores. Each examination has its own scale of raw-score points, but there is only one grading system, whether it be the percentage system with a passing point of 60 or 70 or 75 or a letter system or an honor-point system. If it can be decided what the mean and the standard deviation should be for a given class of students in the marking system, then the procedure to be described will enable one to set up rules or equations for transforming raw scores into marks.

Transformation of Ratings.—The illustration of the transformation procedure will be chosen in the sphere of rating-scale evaluations. Ratings are so often used as criteria of adjustment, either in research or in personnel work, and ratings coming from different sources are so lacking in comparability numerically, that transformation problems in this area are sometimes acute.

Assume that when judge *A* rates 25 individuals for some particular trait on an 11-point scale, he maintains somewhat equal units so far as his own judgments are concerned. Assume that judge *B*, in rating the same 25 individuals for the same trait, also maintains equality of unit. But permit *B*'s unit to differ in size from *A*'s, and permit any particular numerical rating, for example, the rating 7, to mean a higher or lower real value for *B* than it does for *A*. These kinds of discrepancy are probably very common among such ratings. When we average ratings obtained from different judges on the same trait, same individual, we are often averaging things quite different in numerical meaning. If we wish to be relieved of these constant errors, we should transform the ratings into judgments on a common scale before averaging. We may do so by adopting one judge whose ratings seem to cover the scale fairly well as the standard rater and let his mean and standard deviation become the reference values for all distributions.

In Table 19.13 are given ratings assigned to 10 individuals by judge *A* and also by judge *B*. From a much larger sample of ratings by these two judges, we know that *A*'s average rating is 4.08 and that *B*'s average

TABLE 19.13.—PARTIAL LISTS OF RATINGS ASSIGNED BY JUDGES A AND B TO THE SAME INDIVIDUALS FOR THE SAME TRAIT

(1) X_A Ratings by judge A	(2) X_B Ratings by judge B	(3) X_{BA} Ratings by judge B in terms of the mean and sigma of judge A
5	6	3.9
3	7	5.1
2	4	1.5
3	6	3.9
5	8	6.4
7	9	7.6
8	8	6.4
1	4	1.5
7	5	2.7
6	9	7.6
Mean 4.08	6.12	4.07
σ 2.06	1.70	2.08

rating is 6.12. B consistently overrates, apparently, as compared with A. The standard deviation of A's ratings is 2.06 and of B's, 1.70. B's ratings, taken as a whole, cover less range than A's. This may mean that B has not such great discriminating ability as A; or it may mean that he knows people at large who vary more widely than those A knows or that he is more cautious; or there may be other reasons. At any rate, we believe that the individuals rated are really just as variable when B rates them as when A rates them, and they are no higher in the traits rated when B rates them than when A rates them. We really should assume also that A's and B's ratings are equally reliable, or nearly so. We shall now find what B's ratings should be if he had the same general mean and standard deviation as A.

The steps for this procedure are illustrated in Table 19.14.

Step 1. List the ratings used by judge B [column (1)].

Step 2. Find the deviation of each rating from B's mean rating [column (2)].

Step 3. Find the corresponding z-measurements [column (3)].

Step 4. Using A's standard deviation (2.06), multiply every z measurement in column (3) by it. This gives deviations from the mean in terms of A's standard deviation. These are in column (4).

Step 5. Add to each deviation the mean of A's ratings (4.08).

TABLE 19.14.—TRANSLATING RATINGS BY JUDGE *B* INTO TERMS OF THE MEAN AND STANDARD DEVIATION OF JUDGE *A*

(1) X_B Original ratings by judge *B*	(2) x_B Deviation of *B*'s ratings from *B*'s mean	(3) z Standard measurement of *B*'s ratings	(4) x_{BA} *B*'s ratings on *A*'s scale in deviation form	(5) X_{BA} *B*'s ratings in terms of *A*'s mean and σ
9	+2.88	+1.69	+3.48	7.6
8	+1.88	+1.11	+2.29	6.4
7	+0.88	+0.52	+1.07	5.1
6	−0.12	−0.07	−0.14	3.9
5	−1.12	−0.66	−1.36	2.7
4	−2.12	−1.25	−2.57	1.5
3	−3.12	−1.84	−3.78	0.3

We then have the transformed ratings [column (5)]. Now it is seen that when *B* rates an individual 9 on his scale, he means the same as when *A* rates a person 7.6; when *B* rates a person 7, *A* would probably rate him 5.1 (provided, of course, that *A* gave fractional ratings), etc. On the basis of these transformations, *B*'s ratings of the 10 persons in Table 19.13 have been changed, as will be seen in column (3). The mean of this small group of 10 of *B*'s ratings is 4.07, and their sigma is 2.08, both of which are very close to the corresponding values for *A*'s original distribution. In the long run, we should expect them to be exactly the same.

A Transformation Equation.—The transformation procedure just given is longer than need be in practice. Its full length was given for the sake of explaining what actually is going on. It is much more expedient to find a simple equation of transformation in the following manner. In general terms, when the measurements in distribution *B* are to be translated into the scale of distribution *A*, the equation is[1]

$$X_{BA} = \left(\frac{\sigma_A}{\sigma_B}\right) X_B - \left[\left(\frac{\sigma_A}{\sigma_B}\right) M_B - M_A\right] \qquad (19.8)$$

(Equation for transforming a distribution to a different mean and standard deviation)

where X_{BA} = measurement in distribution *B* transformed into the terms of distribution *A*.
X_B = original measurement in distribution *B*.
σ_A = standard deviation of distribution *A*.
σ_B = standard deviation of distribution *B*.
M_B = mean of distribution *B*.
M_A = mean of distribution *A*.

[1] See Appendix A for derivation of equation (19.8).

Formula (19.8) is the same as (12.2) with changed notations. It will also be clear to those who have mastered the discussion of regression equations in Ch. 15, that this is, in effect, a linear regression equation in which it is assumed that $r = +1.00$. The first term on the right of the equation sign is the same as if we wrote the r_{ab} in front of it, but since it is assumed that $r = +1$, we may dispense with it. The second term at the right, within the brackets, is the term a in the ordinary regression equation, except that, again, r_{ab} is not mentioned. We assume a correlation of $+1$, not because we think it actually is $+1$, but because we want the full dispersion of values for X_{ba}. If we put the actual value or r_{ba} in the equation, the σ of the derived values would be r_{ab} times the σ we want.

The ratio σ_a/σ_b appears two times in the equation; so we begin by computing it. In our present problem, the ratio is 2.06/1.70, which equals 1.2118. The two means are known, and when they are substituted in the formula, we have

$$\begin{aligned} X_{BA} &= 1.2118X_B - [(1.2118)(6.12) - 4.08] \\ &= 1.2118X_B - (7.42 - 4.08) \\ &= 1.2118X_B - 3.34 \end{aligned}$$

Rounding,

$$X_{BA} = 1.21X_B - 3.34$$

All we need to do now is to substitute each rating B uses in turn in this equation. Tabulated, this work appears in Table 19.15. The second column gives the product $1.21X_B$ in each case, and the last column comes after deduction of 3.34 in each case. These values coincide with those found by the longer procedure in Table 19.14 with one minor exception which was caused by rounding the coefficient to 1.21.

TABLE 19.15.—TRANSLATING RATINGS BY JUDGE B INTO TERMS OF THE MEAN AND STANDARD DEVIATION OF JUDGE A BY MEANS OF AN EQUATION

The equation: $X_{BA} = 1.12X_B - 3.34$

X_B	$1.21X_B$	X_{BA}
9	10.91	7.6
8	9.69	6.3
7	8.48	5.1
6	7.47	3.9
5	6.06	2.7
4	4.85	1.5
3	3.64	0.3

Fig. 19.7 shows pictorially what happens when the transformation equation is applied. The two original distributions for judges A and B

are shown, each with its own M and σ. By transformation, a rating of 4.4, which is about 1σ below the mean in distribution B, would become a rating of 2, which is about 1σ below the mean in distribution A. A rating of 9.5, which is about 2σ above the mean in distribution B, becomes a rating of 8.1, which is about 2σ above the mean in distribution A. The

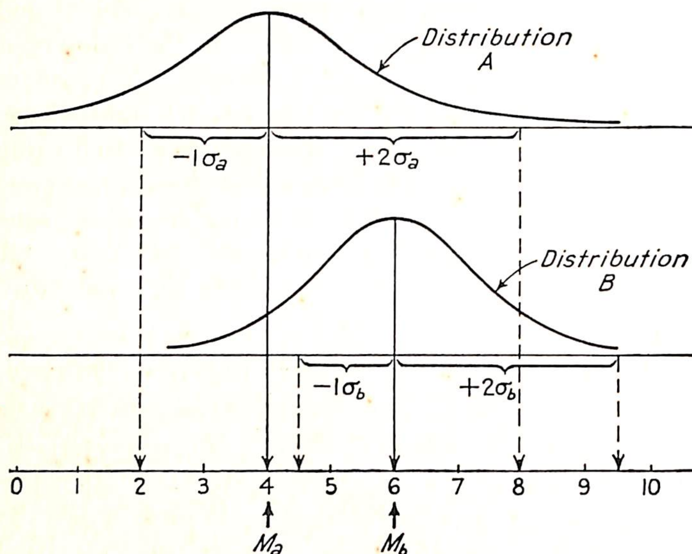


FIG. 19.7.—Illustration of the conversion process represented in the use of formula (19.8) as applied to the data in Table 19.15.

principle is that standard scores or deviates in the two distributions are assumed to be equal psychologically.

Exercises

1. Determine the scale values for difficulty of the test items in Data 19*A* as derived from the two groups. Which group probably has the higher median ability? Explain. Which group is probably more variable in ability? Explain.

2. Determine the scale difficulty for the items in Data 19*B*, both with and without correction for guessing. How is the importance of correcting for guessing shown in your results?

3. Give the corresponding centile positions of the 12 individuals in Data 19*C*. Transform the ranks into *T*-scale values, also into *C*-scale values.

4. Using Data 19*D*, which represent the judgments of 95 members of a well-known symphony orchestra concerning preferences for music of various composers, compute the scale values for the eight composers represented. Make a check on the internal consistency of the data, since it is assumed that dispersions of impressions of single composers are equal.

5. Determine the scale positions in terms of z measurements for the five categories in the ratings of the two words in Data 19*E*. Give the scale position of each word with reference to the indifference category and in terms of the standard deviations of the

two distributions. Are the two standard deviations probably equal? Explain. Are the five categories equally spaced? Explain.

6. Using Data 19*F*, transform the ratings of judges *B* and *C* into the terms of the distribution of judge *A*. Set up equations of transformation. Interpret your results.

DATA 19*A*.—PROPORTIONS OF TWO GROUPS OF STUDENTS PASSING ITEMS IN A COMPLETION EXAMINATION

Item number	1	2	5	7	14	22	84
Group I	.16	.71	.47	.11	.81	.28	.91
Group II	.08	.89	.68	.20	.99	.63	.95

DATA 19*B*.—PROPORTIONS OF STUDENTS PASSING ITEMS IN WHICH THE NUMBER OF ALTERNATIVE RESPONSES VARIES

Number of alternatives	4	4	5	5	3	3	2	2	7	10
Proportion of passes	.47	.87	.30	.98	.36	.84	.99	.75	.82	.29

DATA 19*C*.—SOME RANK ORDERS OF INDIVIDUALS

n Number ranked	*r* Three ranks at random
10	2, 6, 9
25	6, 16, 24
36	1, 12, 30
50	4, 21, 50

DATA 19*D*.—PROPORTIONS OF THE TIMES THAT COMPOSERS INDICATED BY NUMBER AT THE TOPS OF THE COLUMNS WERE PREFERRED TO COMPOSERS AT THE SIDES OF THE ROWS*

Composer	1	2	3	4	5	6	7	8
1. Franck	.500	.242	.842	.916	.884	.589	.411	.126
2. Grieg	.758	.500	.895	.968	.989	.842	.758	.189
3. Haydn	.158	.105	.500	.853	.547	.232	.126	.053
4. Mozart	.084	.032	.147	.500	.189	.126	.053	.032
5. Schubert	.116	.011	.453	.811	.500	.189	.084	.042
6. Tschaikowsky	.411	.158	.768	.874	.811	.500	.326	.116
7. Verdi	.589	.242	.874	.947	.916	.674	.500	.147
8. Herbert	.874	.811	.947	.968	.958	.884	.853	.500

* The number of judges was 95.

These data were abstracted from Folgmann, E. E. E. An experimental study of composer-preferences of four outstanding symphony orchestras. *J. exp. Psychol.*, 1933, **16**, 709–724. By permission of the publisher.

DATA 19*E*.—DISTRIBUTIONS OF JUDGMENTS OF TWO WORDS BY 400 STUDENTS

Word	Very unpleasant	Unpleasant	Indifferent	Pleasant	Very pleasant
Gorgeous..........	1	12	68	215	104
Slang.............	16	119	234	30	1

DATA 19*F*.—RATINGS OF 10 FORMAL DESIGNS BY THREE OBSERVERS

Design	Judge *A*	Judge *B*	Judge *C*
1	9	5	10
2	5	1	5
3	2	3	0
4	7	4	5
5	6	2	3
6	7	7	9
7	4	0	2
8	8	3	6
9	2	4	1
10	6	4	7

APPENDIX A

Some Selected Mathematical Proofs and Derivations

A List of Brief Titles

1. Effect upon a mean of adding a constant
2. Effect upon a mean of multiplying by a constant
3. The mean of a simple linear function
4. Effect upon the standard deviation of adding a constant
5. Effect upon the standard deviation of multiplying by a constant
6. The standard deviation of a simple linear function
7. Variances and standard deviations in combined frequencies
8. Derivation of the formula for the point-biserial r
9. Derivation of the phi coefficient from r_{pbi}
10. Regression coefficients in a two-variable linear equation
11. The mean of a sum of measures
12. The variance and standard deviation in a sum of measures
13. The correlation of sums
14. Linear transformation equation

In this Appendix are presented a few of the derivations or proofs of equations. Selection has been determined by several considerations: (1) Because of their relative simplicity the proofs can be followed by most students. (2) The proofs are illustrative of the manner in which formulas in general are derived. (3) The proofs should help to give insight on some fundamental statistical concepts. (4) The proofs are not commonly found elsewhere. Footnote references in the preceding chapters often indicate sources of derivations of other formulas.

1. *The effect upon a mean of adding a constant to every observed value*

Let X = any observed value in a set of measurements.
C = a constant value added to every X.
M_x = arithmetic mean of all the X values.
$M_{(x+c)}$ = arithmetic mean of all values $(X + C)$.
N = the number of observations in the sample.

Then

$$M_{(x+c)} = \frac{\Sigma(X + C)^*}{N}$$

$$= \frac{\Sigma X}{N} + \frac{NC}{N}$$

$$= M_x + C \qquad \text{(A.1)}$$

* In these equations and those following throughout this Appendix, the summation sign is given without showing the range over which summation is made. Strictly speaking, ΣX should be

In other words, the mean of X values, each augmented by the addition of a constant C, is equal to the mean of the X's plus the same constant. C may have a negative value as well as a positive one.

2. *The effect upon a mean of multiplying each observed value by a constant*

Let M_{CX} = the arithmetic mean of all values $C \times X$, and other symbols be defined as in 1 above.

$$\begin{aligned} M_{cx} &= \frac{\Sigma CX}{N} \\ &= \frac{C\Sigma X}{N} \\ &= CM_x \end{aligned} \tag{A.2}$$

In other words, the mean of X values all multiplied by the same constant is equal to the mean of those values times the constant.

3. *The mean of a linear function of a value*

Let the linear function of X be the regression equation $Y' = a + bX$ (see Ch. 15). We want to find the mean $M_{(a+bX)}$. Here we have a combination of a product of a constant times X, namely, (bX) and also a constant increment (a).

$$\begin{aligned} M_{y'} = M_{(a+bX)} &= \frac{\Sigma(a+bX)}{N} = \frac{Na + b\Sigma X}{N} \\ &= \frac{Na}{N} + \frac{b\Sigma X}{N} \\ &= a + bM_x \end{aligned} \tag{A.3}$$

In other words, the mean of a linear function of X is that same function of the mean of X. This principle is useful in connection with regression equations in general.

4. *Effect upon the standard deviation of adding a constant to each observed value*

Using the same symbols as above, with the addition of:

σ_x = standard deviation of the X values.
$\sigma_{(x+c)}$ = standard deviation of all values $(X + C)$.
x = a deviation of X from M_x.
$x_{(x+c)}$ = deviation of $(X + C)$ from the mean $(M_x + C)$.

We find that

$$\begin{aligned} x_{(x+c)} &= (X + C) - (M_x + C) \\ &= X - M_x \\ &= x \end{aligned}$$

---

written here as

$$\sum_{1}^{N} X$$

to show that the N values of the sample are included. The omission makes for easier reading, particularly where formulas become complicated. It is believed that in all instances the range of summation will be clear; if not directly from the formula, at least from the context.

From this it follows that

$$\Sigma x^2_{(x+c)} = \Sigma x^2$$
$$\sigma^2_{(x+c)} = \sigma^2_x$$

and

$$\sigma_{(x+c)} = \sigma_x \qquad \text{(A.4)}$$

In other words, adding a constant to every observed value has no effect upon the standard deviation.

5. *Effect upon the standard deviation of multiplying each observed value by a constant, C*

Let σ_{cx} = standard deviation of the products CX.
From (A.2) above,

$$M_{cx} = CM_x$$

Therefore,

$$\begin{aligned} x_{cx} &= CX - CM_x \\ &= C(X - M_x) \\ &= Cx \\ \sigma^2_{cx} &= \frac{C^2\Sigma x^2}{N} \\ &= C^2\sigma^2_x \end{aligned} \qquad \text{(A.5)}$$

Taking square roots of both sides of (A.5)

$$\sigma_{cx} = C\sigma_x \qquad \text{(A.6)}$$

6. *Standard deviation of a linear function of X*

If the function of X is $a + bX$, the mean of this function, from (A.3) above, is equal to $a + bM_x$. Each deviation of this function (Y) from its mean is, therefore,

$$\begin{aligned} y_{(a+bX)} &= (a + bX) - (a + bM_x) \\ &= bX - bM_x \\ &= b(X - M_x) \\ &= bx \end{aligned}$$

From (A.6), we deduce that $\sigma_{bx} = b\sigma_x$. Therefore,

$$\sigma_{(a+bx)} = b\sigma_x \qquad \text{(A.7)}$$

Thus, wherever we use a simple regression equation of the form $Y' = a + bX$, the standard deviation of Y' equals $b\sigma_x$.

7. *Variances and standard deviations of combined distributions*

Assume two sample distributions A and B, whose frequencies are summed to form a total distribution T.

Let M_a, M_b, and M_t = means of distributions A, B, and T, respectively.
n_a, n_b, and N = numbers of cases in corresponding distributions.
X_a, X_b, and X_t = measures in the three distributions, respectively.
x_a, x_b, and x_t = deviations of measures from the means of their respective distributions.

x_{at} and x_{bt} = deviations of measures in distributions A and B, respectively, from M_t.

d_a and d_b = deviations of means of distributions A and B, respectively, from M_t.

From the preceding,

$$d_a = M_a - M_t, \qquad \text{and} \qquad d_b = M_b - M_t \tag{A.8}$$

Transposing,

$$M_t = M_a - d_a, \qquad \text{and} \qquad M_t = M_b - d_b \tag{A.9}$$

By definition given above, and from (A.9) and (A.8),

$$x_{at} = X_a - M_t = X_a - M_a + d_a = x_a + d_a$$

and

$$x_{bt} = X_b - M_t = X_b - M_b + d_b = x_b + d_b$$

Squaring both sides of these equations,

$$x^2_{at} = (x_a + d_a)^2 = x^2_a + d^2_a + 2x_a d_a$$

and

$$x^2_{bt} = (x_b + d_b)^2 = x^2_b + d^2_b + 2x_b d_b$$

Summing for all measures in either distribution,

$$\Sigma x^2_{at} = \Sigma x^2_a + n_a d^2_a + 2d_a \Sigma x_a$$

and

$$\Sigma x^2_{bt} = \Sigma x^2_b + n_b d^2_b + 2d_b \Sigma x_b$$

Now both Σx_a and Σx_b equal zero, which eliminates the last terms from the last two equations. The sum of squares in the total distribution is the combination of Σx^2_{at} and Σx^2_{bt}, in other words,

$$\Sigma x^2_t = \Sigma x^2_a + n_a d^2_a + \Sigma x^2_b + n_b d^2_b \tag{A.10a}$$

Or, by combining terms,

$$\Sigma x^2_t = (\Sigma x^2_a + \Sigma x^2_b) + (n_a d^2_a n_b d^2_b) \tag{A.10b}$$

This proof has involved the combination of only two sample distributions. It can readily be generalized to include any number of samples, by adding, by analogy, additional equations in each step taken above. Equation (A.10*b*) is identical with equation (5.17) in Ch. 5.

8. *Formula for the point-biserial coefficient of correlation,* r_{pbi}

Let X be a continuous variable, continuously measured.

Y be a genuine dichotomy, with point values of 0 and +1.

The cases in the favored category have values of +1.

N = total number of cases.

N_p = number of cases in the favored category ($N_p = pN$).

N_q = number of cases in the other category ($N_q = qN$. $N_p + N_q = N$).

M_x = arithmetic mean of the X values.

σ_x = standard deviation of the X values.

M_p = mean of the X values in the favored category on Y.

M_q = mean of the X values for the remaining category.

p = the proportion of the cases in the favored category ($p = N_p/N$).
$q = 1 - p$. q also equals N_q/N.
M_y = the mean of the point values in variable Y. It can be shown to equal p (see Table 9.3).
σ_y = the standard deviation in the point values. It can be shown to equal $\sqrt{pq}$ (see Table 9.3).

The point-biserial r is a product-moment correlation coefficient. There are several ways of deriving the formula for r_{pbi}. Let us start with the basic formula for the Pearson r,

$$r_{yx} = \frac{\Sigma xy}{N\sigma_x\sigma_y} \tag{A.11}$$

where $x = X - M_x$.
$y = Y - M_y$.

Therefore,

$$\begin{aligned} \Sigma xy &= \Sigma(X - M_x)(Y - M_y) \\ &= \Sigma XY - M_y\Sigma X - M_x\Sigma Y + NM_xM_y \end{aligned} \tag{A.12}$$

Substituting NM_x for ΣX and NM_y for ΣY in (A.12),

$$\begin{aligned} \Sigma xy &= \Sigma XY - NM_xM_y - NM_xM_y + NM_xM_y \\ &= \Sigma XY - NM_xM_y \end{aligned} \tag{A.13}$$

Substituting (A.13) in (A.11),

$$r_{yx} = \frac{\Sigma XY - NM_xM_y}{N\sigma_x\sigma_y} \tag{A.14}$$

Making some other substitutions:

$$\Sigma XY = N_pM_p,\quad NM_xM_y = NM_xp = N_pM_x, \quad \text{and} \quad \sigma_y = \sqrt{pq},$$

We get

$$r_{yx} = \frac{N_pM_p - N_pM_x}{N\sigma_x\sqrt{pq}} \tag{A.15}$$

Dividing numerator and denominator of (A.15) by N,

$$r_{yx} = \frac{pM_p - pM_x}{\sigma_x\sqrt{pq}} = \frac{(M_p - M_x)p}{\sigma_x\sqrt{pq}} \tag{A.16}$$

Dividing numerator and denominator of (A.16) by $\sqrt{p}$,

$$r_{yx} = \frac{(M_p - M_x)}{\sigma_x}\sqrt{\frac{p}{q}} \tag{A.17}$$

This is one form of the equation for the point-biserial r. If we want the form involving M_q rather than M_x, some further proof is required.

$$M_x = pM_p + qM_q$$

So that

$$\begin{aligned} M_p - M_x &= M_p - pM_p - qM_q \\ &= (1 - p)M_p - qM_q \\ &= qM_p - qM_q \\ &= q(M_p - M_q) \end{aligned} \tag{A.18}$$

Substituting (A.18) in (A.17),

$$r_{pbi} = \frac{(M_p - M_q)\sqrt{pq}}{\sigma_x} \tag{A.19}$$

9. *Derivation of the formula for* ϕ *from* r_{pbi}

Phi is a product-moment correlation in a 2 × 2 contingency table where both variables are genuine dichotomies and the distributions are point distributions, with values of +1 and 0. Let the symbols used be defined in the two following tables, one based upon frequencies and the other upon corresponding proportions.

FREQUENCIES

	+1	0	Both
+1	a	b	N_p
0	c	d	N_q
Both	$N_{p'}$	$N_{q'}$	N

PROPORTIONS

	+1	0	Both
+1	α	β	p
0	γ	δ	q
Both	p'	q'	1.00

In these point distributions,

$$M_p = \frac{a}{N_p} = \frac{\alpha}{p}$$

$$M_q = \frac{c}{N_q} = \frac{\gamma}{q}$$

$$\sigma_x = \sqrt{p'q'}$$

Substituting these values in (A.19), we have

$$r = \phi = \frac{\left(\frac{\alpha}{p} - \frac{\gamma}{q}\right)\sqrt{pq}}{\sqrt{p'q'}} \tag{A.20}$$

Now

$$\frac{\alpha}{p} - \frac{\gamma}{q} = \frac{\alpha q - \gamma p}{pq} \tag{A.21}$$

And since $p = \alpha + \beta$ and $q = \gamma + \delta$, the right side of (A.21) becomes

$$\frac{\alpha(\gamma + \delta) - \gamma(\alpha + \beta)}{pq} = \frac{\alpha\gamma + \alpha\delta - \alpha\gamma - \beta\gamma}{pq} = \frac{\alpha\delta - \beta\gamma}{pq} \tag{A.22}$$

Substituting (A.22) in (A.20),

$$\phi = \frac{(\alpha\delta - \beta\gamma)\sqrt{pq}}{pq\sqrt{p'q'}}$$

$$\phi = \frac{\alpha\delta - \beta\gamma}{\sqrt{pqp'q'}} \tag{A.23}$$

10. *Regression coefficients in a two-variable linear equation*

Let the general regression equation for a straight line be

$$Y' = a + bX$$

Problem: To find for any set of data involving corresponding X and Y those values of a and b which will make $\Sigma(Y - Y')^2$ a minimum.

We first set up an equation involving the expression $(Y - Y')$:

$$(Y - Y') = Y - a - bX$$

Squaring both sides we have an expression for the discrepancy squared:

$$\begin{aligned}(Y - Y')^2 &= (Y - a - bX)^2 \\ &= Y^2 + a^2 + b^2X^2 - 2aY - 2bXY + 2abX\end{aligned}$$

Summing for all observations,

$$\Sigma(Y - Y')^2 = \Sigma Y^2 + Na^2 + b^2\Sigma X^2 - 2a\Sigma Y - 2b\Sigma XY + 2ab\Sigma X \qquad (A.24)$$

The partial derivatives of (A.24) are

$$\frac{\partial[\Sigma(Y - Y')^2]}{\partial a} = 2Na - 2\Sigma Y + 2b\Sigma X \qquad (A.25)$$

$$\frac{\partial[\Sigma(Y - Y')^2]}{\partial b} = 2b\Sigma X^2 - 2\Sigma XY + 2a\Sigma X \qquad (A.26)$$

Setting derivative (A.25) equal to zero, we have

$$2Na - 2\Sigma Y + 2b\Sigma X = 0$$

Or

$$Na - \Sigma Y + b\Sigma X = 0$$

Transposing, we have

$$Na + b\Sigma X = \Sigma Y \qquad (A.27)$$

Setting derivative (A.26) equal to zero, we have

$$2b\Sigma X^2 - 2\Sigma XY + 2a\Sigma X = 0$$

Or

$$b\Sigma X^2 - \Sigma XY + a\Sigma X = 0$$

Transposing, we have

$$a\Sigma X + b\Sigma X^2 = \Sigma XY \qquad (A.28)$$

(A.27) and (A.28) provide us with two *normal equations* which, solved simultaneously, give us formulas for deriving a and b from the observations X and Y.

Dividing (A.27) by N, we have

$$a + \frac{(\Sigma X)b}{N} = \frac{\Sigma Y}{N}$$

$$a + M_xb = M_y$$

Transposing,

$$a = M_y - M_xb \qquad (A.29)$$

Substituting (A.29) in (A.28) we have

$$(\Sigma X)M_y - (\Sigma X)M_xb + (\Sigma X^2)b = \Sigma XY$$

Collecting terms and transposing,

$$[(\Sigma X^2) - (\Sigma X)M_x]b = \Sigma XY - (\Sigma X)M_y$$

Solving for b,

$$b = \frac{\Sigma XY - (\Sigma X)M_y}{(\Sigma X^2) - (\Sigma X)M_x} \qquad \text{(A.30)}$$

Multiplying numerator and denominator by N,

$$b = \frac{N\Sigma XY - (\Sigma X)(\Sigma Y)}{N\Sigma X^2 - (\Sigma X)^2} \qquad \text{(A.31)}$$

11. *The mean of a sum of measurements*

a. For equally weighted measurements.

Let X_1 and X_2 be two independently derived measures of the same individual. Let X_1 and X_2 be summed for each individual, giving a composite measure $X_1 + X_2$. The problem is to find the mean of the composite, $M_{(x_1+x_2)}$.

$$\begin{aligned} M_{(x_1+x_2)} &= \frac{\Sigma(X_1 + X_2)}{N} \\ &= \frac{\Sigma X_1 + \Sigma X_2}{N} \\ &= \frac{\Sigma X_1}{N} + \frac{\Sigma X_2}{N} \\ &= M_1 + M_2 \end{aligned} \qquad \text{(A.32)}$$

where M_1 = mean of X_1 values.

M_2 = mean of X_2 values.

For the general case, in which there are n measurements of each individual, it can be similarly shown that

$$M_{(x_1+x_2+\cdots+x_n)} = M_1 + M_2 + \cdots + M_n \qquad \text{(A.33)}$$

If we let the symbols M_s = mean of an unweighted sum of n measures and M_i = the mean of any one of the measures X_1 to X_n inclusive, we may write equation (A.33) in more economical form as

$$M_s = \Sigma M_i \qquad \text{(A.34)}$$

In other words, when measures are summed without weighting, the mean of the sums is equal to the sum of the means.

b. For differentially weighted measurements.

When the measurements X_1 and X_2 are weighted by multipliers w_1 and w_2, respectively,

$$\begin{aligned} M_{(w_1x_1+w_2x_2)} &= \frac{\Sigma(w_1X_1 + w_2X_2)}{N} \\ &= \frac{w_1\Sigma X_1 + w_2\Sigma X_2}{N} \\ &= \frac{w_1\Sigma X_1}{N} + \frac{w_2\Sigma X_2}{N} \\ &= w_1M_1 + w_2M_2 \end{aligned}$$

To describe the general case, with n measurements,

$$M_{(w_1x_1+w_2x_2+\cdots+w_nx_n)} = w_1M_1 + w_2M_2 + \cdots + w_nM_n \qquad \text{(A.35)}$$

If M_{ws} symbolizes the mean of a weighted composite, and M_i symbolizes the mean of any one measurement that enters into it, we may write equation (A.35) in abbreviated form:

$$M_{ws} = \Sigma w_i M_i \tag{A.36}$$

12. *Variance and standard deviation of a sum*

a. When measurements are equally weighted.

Let X_1 and X_2 be two independently derived measures of the same individual, summed without weighting to obtain a composite measure. The variance of the composite measures is given by the equation

$$\sigma^2_{(x_1+x_2)} = \frac{\Sigma(x_1 + x_2)^2}{N} \tag{A.37}$$

where $(x_1 + x_2)$ = a deviation of $(X_1 + X_2)$ from $M_{(x_1+x_2)}$.* Expanding the binomial in (A.37),

$$\sigma^2_{(x_1+x_2)} = \frac{\Sigma(x^2_1 + x^2_2 + 2x_1x_2)}{N}$$
$$= \frac{\Sigma x^2_1}{N} + \frac{\Sigma x^2_2}{N} + 2\frac{\Sigma x_1x_2}{N} \tag{A.38}$$

The most meaningful interpretation to make of (A.38) in this development is to say that the first term on the right of the equality sign is the variance in X_1, the second term is the variance in X_2, and the third term is twice the covariance between X_1 and X_2. It will be helpful, next, to relate the covariance term to the correlation between X_1 and X_2. By the Pearson product-moment formula,

$$r_{12} = \frac{\Sigma x_1x_2}{N\sigma_1\sigma_2} \tag{A.39}$$

Multiplying both sides of (A.39) by $\sigma_1\sigma_2$,

$$r_{12}\sigma_1\sigma_2 = \frac{\Sigma x_1x_2}{N} \tag{A.40}$$

Substituting σ^2_1, σ^2_2, and $r_{12}\sigma_1\sigma_2$ in (A.38), we have

$$\sigma^2_{(x_1+x_2)} = \sigma^2_1 + \sigma^2_2 + 2r_{12}\sigma_1\sigma_2 \tag{A.41}$$

Taking square roots of both sides of (A.41),

$$\sigma_{(x_1+x_2)} = \sqrt{\sigma^2_1 + \sigma^2_2 + 2r_{12}\sigma_1\sigma_2} \tag{A.42}$$

In other words, the variance of an unweighted sum of two measures is equal to the sum of the variances of the components plus two times their covariance. To generalize to any number of unweighted components, and remembering that we will have as many covariance terms as there are *pairs* of components,

$$\sigma^2_{(x_1+x_2+\cdots+x_n)} = \sigma^2_1 + \sigma^2_2 + \cdots + \sigma^2_n + 2r_{12}\sigma_1\sigma_2 + 2r_{13}\sigma_1\sigma_3 + \cdots + 2r_{1n}\sigma_1\sigma_n + \cdots + 2r_{(n-1)n}\sigma_{(n-1)}\sigma_n$$

Let σ^2_s = the variance of an unweighted sum of any number of measures.
σ^2_i = variance of any measure from 1 to n, inclusive.

* The deviation of a composite of two values from the mean of the composite equals $x_1 + x_2$, for

$$(X_1 + X_2) - (M_1 + M_2) = (X_1 - M_1) + (X_2 - M_2) = x_1 + x_2$$

Then

$$\sigma^2_s = \Sigma\sigma^2_i + 2\Sigma r_{ij}\sigma_i\sigma_j \quad \text{(where } i < j\text{)} \tag{A.43}$$

By square roots, the standard deviation of a sum is given by

$$\sigma_s = \sqrt{\Sigma\sigma^2_i + 2\Sigma r_{ij}\sigma_i\sigma_j} \quad \text{(where } i < j\text{)} \tag{A.44}$$

b. When measurements are differentially weighted.

Let the weights to be applied to $X_1, X_2, \ldots, X_n$ be $w_1, w_2, \ldots, w_n$, respectively. For the variance of the sum of two weighted measurements:

$$\begin{aligned}\sigma^2_{(w_1x_1+w_2x_2)} &= \frac{\Sigma(w_1x_1 + w_2x_2)^2}{N} \\ &= \frac{\Sigma(w^2_1x^2_1 + w^2_2x^2_2 + 2w_1w_2x_1x_2)}{N} \\ &= \frac{w^2_1\Sigma x^2_1}{N} + \frac{w^2_2\Sigma x^2_2}{N} + 2w_1w_2\frac{\Sigma x_1x_2}{N}\end{aligned}$$

Making substitutions similar to those made in (A.38),

$$\sigma^2_{(w_1x_1+w_2x_2)} = w^2_1\sigma^2_1 + w^2_2\sigma^2_2 + 2r_{12}w_1w_2\sigma_1\sigma_2 \tag{A.45}$$

In other words, the variance of a weighted sum of two measures equals the sum of the component variances, each weighted by its weight squared, plus twice the covariance multiplied by the product of the weights. The standard deviation, by taking square roots, is

$$\sigma_{(w_1x_1+w_2x_2)} = \sqrt{w^2_1\sigma^2_1 + w^2_2\sigma^2_2 + 2r_{12}w_1w_2\sigma_1\sigma_2} \tag{A.46}$$

Generalized to include n components and to apply the symbols as defined in (A.43),

$$\sigma_{ws} = \sqrt{\Sigma w^2_i\sigma^2_i + 2\Sigma r_{ij}w_iw_j\sigma_i\sigma_j} \quad \text{(where } i < j\text{)} \tag{A.47}$$

13. *Correlation of sums*

a. Correlation between one variable, C, and an unweighted sum of two other variables, X_1 and X_2.

Applying the Pearson product-moment formula to this problem,

$$\begin{aligned}r_{c(x_1+x_2)} &= \frac{\Sigma c(x_1 + x_2)}{N\sigma_c\sigma_{(x_1+x_2)}} \\ &= \frac{\Sigma cx_1 + \Sigma cx_2}{N\sigma_c\sigma_{(x_1+x_2)}}\end{aligned} \tag{A.48}$$

Now $\Sigma cx_1 = Nr_{c1}\sigma_c\sigma_1$ and $\Sigma cx_2 = Nr_{c2}\sigma_c\sigma_2$

Substituting these values in (A.48), we have

$$r_{c(x_1+x_2)} = \frac{Nr_{c1}\sigma_c\sigma_1 + Nr_{c2}\sigma_c\sigma_2}{N\sigma_c\sigma_{(x_1+x_2)}}$$

Eliminating $N\sigma_c$, and expanding the standard deviation of the sum,

$$r_{c(x_1+x_2)} = \frac{r_{c1}\sigma_1 + r_{c2}\sigma_2}{\sqrt{\sigma_1^2 + \sigma^2_2 + 2r_{12}\sigma_1\sigma_2}} \tag{A.49}$$

Let r_{cs} = correlation of the sum of n unweighted measures with C.
X_i = any variable from 1 to n, inclusive.
r_{ci} = correlation of C with any variable 1 to n.
X_j = any variable with a greater subscript number than X_i.
Extended to the general case, (A.49) becomes

$$r_{cs} = \frac{\Sigma r_{ci}\sigma_i}{\sqrt{\Sigma\sigma^2_i + 2\Sigma r_{ij}\sigma_i\sigma_j}} \qquad \text{(where } i < j\text{)} \qquad \text{(A.50)}$$

b. Correlation of one variable, C, with the sum of differentially weighted variables.
Let $w_1, w_2, \ldots, w_n$ weights applied to measures $X_1, X_2, \ldots, X_n$, respectively.
For the sum of two variables, by Pearson's formula,

$$r_{c(w_1x_1+w_2x_2)} = \frac{\Sigma c(w_1x_1 + w_2x_2)}{N\sigma_c\sigma_{(w_1x_1+w_2x_2)}}$$
$$= \frac{w_1\Sigma cx_1 + w_2\Sigma cx_2}{N\sigma_c\sigma_{(w_1x_1+w_2x_2)}}$$

Making substitutions as in (A.48) above,

$$r_{c(w_1x_1+w_2x_2)} = \frac{Nw_1r_{c1}\sigma_c\sigma_1 + Nw_2r_{c2}\sigma_c\sigma_2}{N\sigma_c\sigma_{(w_1x_1+w_2x_2)}}$$

Eliminating $N\sigma_c$ and expanding the standard deviation of the weighted sum,

$$r_{c(w_1x_1+w_2x_2)} = \frac{w_1r_{c1}\sigma_1 + w_2r_{c2}\sigma_2}{\sqrt{w^2{}_1\sigma^2{}_1 + w^2{}_2\sigma^2{}_2 + 2\Sigma r_{12}w_1w_2\sigma_1\sigma_2}} \qquad \text{(A.51)}$$

Generalizing to any number of weighted components,

$$r_{c(ws)} = \frac{\Sigma w_ir_{ci}\sigma_i}{\sqrt{\Sigma w^2{}_i\sigma^2{}_i + 2\Sigma r_{ij}w_iw_j\sigma_i\sigma_j}} \qquad \text{(where } i < j\text{)} \qquad \text{(A.52)}$$

c. Correlation of two unweighted composites.

Without presenting the proof, which is quite analogous to those just presented, two formulas will be given here for the correlation of two composite measures from information about correlations among the components.

Let X_i and X_j be any two measures in the first composite, C_1,
X_u and X_v be any two measures in the second composite, C_2.

By analogy to (A.50) and (A.52), the following equations apply. (A.53) is for two unweighted composites, and (A.54) for weighted composites. (A.54) reduces to (A.53) if all weights are +1.

$$r_{c_1c_2} = \frac{\Sigma(\sigma_i\Sigma r_{iu}\sigma_u)}{\sqrt{\Sigma\sigma^2{}_i + 2\Sigma r_{ij}\sigma_i\sigma_j}\sqrt{\Sigma\sigma^2{}_u + 2\Sigma r_{uv}\sigma_u\sigma_v}} \qquad \text{(where } i < j \text{ and } u < v\text{)} \qquad \text{(A.53)}$$

$$r_{wc_1wc_2} = \frac{\Sigma(w_i\sigma_i\Sigma r_{iu}w_u\sigma_u)}{\sqrt{\Sigma w^2{}_i\sigma^2{}_i + 2\Sigma r_{ij}w_i\sigma_iw_j\sigma_j}\sqrt{\Sigma w^2{}_u\sigma^2{}_u + 2\Sigma r_{uv}w_u\sigma_uw_v\sigma_v}}$$
$$\text{(where } i < j \text{ and } u < v\text{)} \qquad \text{(A.54)}$$

14. *Linear transformation of values in one distribution to corresponding standard-score positions in another*

Problem: Given a distribution of observed values, to find a linear equation which will determine for each value one that deviates as much in terms of standard-deviation units from the mean in another distribution of similar values and in the same direction.

Let X_a = a value in distribution *A*.
M_a = mean of values in distribution *A*.
σ_a = standard deviation in distribution *A*.
X_b = a value in distribution *B*.
M_b = mean of values in distribution *B*.
σ_b = standard deviation in distribution *B*.
X_{ba} = a value in distribution *A* equivalent to one in distribution *B*, where equivalence is as defined above.

Assume, as the problem statement requires, that standard measures or deviates in the two distributions are equal. In equation form,

$$\frac{X_{ba} - M_a}{\sigma_a} = \frac{X_b - M_b}{\sigma_b} \tag{A.55}$$

Multiplying (A.55) by σ_a,

$$X_{ba} - M_a = \frac{X_b\sigma_a - M_b\sigma_a}{\sigma_b}$$
$$= \left(\frac{\sigma_a}{\sigma_b}\right) X_b - \left(\frac{\sigma_a}{\sigma_b}\right) M_b$$

Transposing,

$$X_{ba} = \left(\frac{\sigma_a}{\sigma_b}\right) X_b - \left(\frac{\sigma_a}{\sigma_b}\right) M_b + M_a$$
$$= \left(\frac{\sigma_a}{\sigma_b}\right) X_b - \left[\left(\frac{\sigma_a}{\sigma_b}\right) M_b - M_a\right] \tag{A.56}$$

APPENDIX B

TABLES

A List of Brief Titles

A. Squares and square roots of numbers 1 to 1,000
B. Proportions of area under the normal distribution curve
C. Standard scores and ordinates corresponding to areas under the normal curve
D. Significant coefficients of correlation and t ratios
E. Chi square
F. F ratio
G. Functions of p, q, z, and y
H. Fisher's **z** for different values of r
J. Trigonometric functions
K. Four-place logarithms of numbers

TABLE A.—SQUARES AND SQUARE ROOTS OF NUMBERS FROM 1 TO 1,000*

Number	Square	Square root	Number	Square	Square root
1	1	1.0000	41	16 81	6.4031
2	4	1.4142	42	17 64	6.4807
3	9	1.7321	43	18 49	6.5574
4	16	2.0000	44	19 36	6.6332
5	25	2.2361	45	20 25	6.7082
6	36	2.4495	46	21 16	6.7823
7	49	2.6458	47	22 09	6.8557
8	64	2.8284	48	23 04	6.9282
9	81	3.0000	49	24 01	7.0000
10	1 00	3.1623	50	25 00	7.0711
11	1 21	3.3166	51	26 01	7.1414
12	1 44	3.4641	52	27 04	7.2111
13	1 69	3.6056	53	28 09	7.2801
14	1 96	3.7417	54	29 16	7.3485
15	2 25	3.8730	55	30 25	7.4162
16	2 56	4.0000	56	31 36	7.4833
17	2 89	4.1231	57	32 49	7.5498
18	3 24	4.2426	58	33 64	7.6158
19	3 61	4.3589	59	34 81	7.6811
20	4 00	4.4721	60	36 00	7.7460
21	4 41	4.5826	61	37 21	7.8102
22	4 84	4.6904	62	38 44	7.8740
23	5 29	4.7958	63	39 69	7.9373
24	5 76	4.8990	64	40 96	8.0000
25	6 25	5.0000	65	42 25	8.0623
26	6 76	5.0990	66	43 56	8.1240
27	7 29	5.1962	67	44 89	8.1854
28	7 84	5.2915	68	46 24	8.2462
29	8 41	5.3852	69	47 61	8.3066
30	9 00	5.4772	70	49 00	8.3666
31	9 61	5.5678	71	50 41	8.4261
32	10 24	5.6569	72	51 84	8.4853
33	10 89	5.7446	73	53 29	8.5440
34	11 56	5.8310	74	54 76	8.6023
35	12 25	5.9161	75	56 25	8.6603
36	12 96	6.0000	76	57 76	8.7178
37	13 69	6.0828	77	59 29	8.7750
38	14 44	6.1644	78	60 84	8.8318
39	15 21	6.2450	79	62 41	8.8882
40	16 00	6.3246	80	64 00	8.9443

* From Sorenson. Statistics for students of psychology and education.

TABLE A.—SQUARES AND SQUARE ROOTS OF NUMBERS FROM 1 TO 1,000.*—(*Continued*)

Number	Square	Square root	Number	Square	Square root
81	65 61	9.0000	121	1 46 41	11.0000
82	67 24	9.0554	122	1 48 84	11.0454
83	68 89	9.1104	123	1 51 29	11.0905
84	70 56	9.1652	124	1 53 76	11.1355
85	72 25	9.2195	125	1 56 25	11.1803
86	73 96	9.2736	126	1 58 76	11.2250
87	75 69	9.3274	127	1 61 29	11.2694
88	77 44	9.3808	128	1 63 84	11.3137
89	79 21	9.4340	129	1 66 41	11.3578
90	81 00	9.4868	130	1 69 00	11.4018
91	82 81	9.5394	131	1 71 61	11.4455
92	84 64	9.5917	132	1 74 24	11.4891
93	86 49	9.6437	133	1 76 89	11.5326
94	88 36	9.6954	134	1 79 56	11.5758
95	90 25	9.7468	135	1 82 25	11.6190
96	92 16	9.7980	136	1 84 96	11.6619
97	94 09	9.8489	137	1 87 69	11.7047
98	96 04	9.8995	138	1 90 44	11.7473
99	98 01	9.9499	139	1 93 21	11.7898
100	1 00 00	10.0000	140	1 96 00	11.8322
101	1 02 01	10.0499	141	1 98 81	11.8743
102	1 04 04	10.0995	142	2 01 64	11.9164
103	1 06 09	10.1489	143	2 04 49	11.9583
104	1 08 16	10.1980	144	2 07 36	12.0000
105	1 10 25	10.2470	145	2 10 25	12.0416
106	1 12 36	10.2956	146	2 13 16	12.0830
107	1 14 49	10.3441	147	2 16 09	12.1244
108	1 16 64	10.3923	148	2 19 04	12.1655
109	1 18 81	10.4403	149	2 22 01	12.2066
110	1 21 00	10.4881	150	2 25 00	12.2474
111	1 23 21	10.5357	151	2 28 01	12.2882
112	1 25 44	10.5830	152	2 31 04	12.3288
113	1 27 69	10.6301	153	2 34 09	12.3693
114	1 29 96	10.6771	154	2 37 16	12.4097
115	1 32 25	10.7238	155	2 40 25	12.4499
116	1 34 56	10.7703	156	2 43 36	12.4900
117	1 36 89	10.8167	157	2 46 49	12.5300
118	1 39 24	10.8628	158	2 49 64	12.5698
119	1 41 61	10.9087	159	2 52 81	12.6095
120	1 44 00	10.9545	160	2 56 00	12.6491

* From Sorenson. Statistics for students of psychology and education.

TABLE A.—SQUARES AND SQUARE ROOTS OF NUMBERS FROM 1 TO 1,000.*—(*Continued*)

Number	Square	Square root	Number	Square	Square root
161	2 59 21	12.6886	201	4 04 01	14.1774
162	2 62 44	12.7279	202	4 08 04	14.2127
163	2 65 69	12.7671	203	4 12 09	14.2478
164	2 68 96	12.8062	204	4 16 16	14.2829
165	2 72 25	12.8452	205	4 20 25	14.3178
166	2 75 56	12.8841	206	4 24 36	14.3527
167	2 78 89	12.9228	207	4 28 49	14.3875
168	2 82 24	12.9615	208	4 32 64	14.4222
169	2 85 61	13.0000	209	4 36 81	14.4568
170	2 89 00	13.0384	210	4 41 00	14.4914
171	2 92 41	13.0767	211	4 45 21	14.5258
172	2 95 84	13.1149	212	4 49 44	14.5602
173	2 99 29	13.1529	213	4 53 69	14.5945
174	3 02 76	13.1909	214	4 57 96	14.6287
175	3 06 25	13.2288	215	4 62 25	14.6629
176	3 09 76	13.2665	216	4 66 56	14.6969
177	3 13 29	13.3041	217	4 70 89	14.7309
178	3 16 84	13.3417	218	4 75 24	14.7648
179	3 20 41	13.3791	219	4 79 61	14.7986
180	3 24 00	13.4164	220	4 84 00	14.8324
181	3 27 61	13.4536	221	4 88 41	14.8661
182	3 31 24	13.4907	222	4 92 84	14.8997
183	3 34 89	13.5277	223	4 97 29	14.9332
184	3 38 56	13.5647	224	5 01 76	14.9666
185	3 42 25	13.6015	225	5 06 25	15.0000
186	3 45 96	13.6382	226	5 10 76	15.0333
187	3 49 69	13.6748	227	5 15 29	15.0665
188	3 53 44	13.7113	228	5 19 84	15.0997
189	3 57 21	13.7477	229	5 24 41	15.1327
190	3 61 00	13.7840	230	5 29 00	15.1658
191	3 64 81	13.8203	231	5 33 61	15.1987
192	3 68 64	13.8564	232	5 38 24	15.2315
193	3 72 49	13.8924	233	5 42 89	15.2643
194	3 76 36	13.9284	234	5 47 56	15.2971
195	3 80 25	13.9642	235	5 52 25	15.3297
196	3 84 16	14.0000	236	5 56 96	15.3623
197	3 88 09	14.0357	237	5 61 69	15.3948
198	3 92 04	14.0712	238	5 66 44	15.4272
199	3 96 01	14.1067	239	5 71 21	15.4596
200	4 00 00	14.1421	240	5 76 00	15.4919

* From Sorenson. Statistics for students of psychology and education.

TABLE A.—SQUARES AND SQUARE ROOTS OF NUMBERS FROM 1 TO 1,000.*—(*Continued*)

Number	Square	Square root	Number	Square	Square root
241	5 80 81	15.5242	281	7 89 61	16.7631
242	5 85 64	15.5563	282	7 95 24	16.7929
243	5 90 49	15.5885	283	8 00 89	16.8226
244	5 95 36	15.6205	284	8 06 56	16.8523
245	6 00 25	15.6525	285	8 12 25	16.8819
246	6 05 16	15.6844	286	8 17 96	16.9115
247	6 10 09	15.7162	287	8 23 69	16.9411
248	6 15 04	15.7480	288	8 29 44	16.9706
249	6 20 01	15.7797	289	8 35 21	17.0000
250	6 25 00	15.8114	290	8 41 00	17.0294
251	6 30 01	15.8430	291	8 46 81	17.0587
252	6 35 04	15.8745	292	8 52 64	17.0880
253	6 40 09	15.9060	293	8 58 49	17.1172
254	6 45 16	15.9374	294	8 64 36	17.1464
255	6 50 25	15.9687	295	8 70 25	17.1756
256	6 55 36	16.0000	296	8 76 16	17.2047
257	6 60 49	16.0312	297	8 82 09	17.2337
258	6 65 64	16.0624	298	8 88 04	17.2627
259	6 70 81	16.0935	299	8 94 01	17.2916
260	6 76 00	16.1245	300	9 00 00	17.3205
261	6 81 21	16.1555	301	9 06 01	17.3494
262	6 86 44	16.1864	302	9 12 04	17.3781
263	6 91 69	16.2173	303	9 18 09	17.4069
264	6 96 96	16.2481	304	9 24 16	17.4356
265	7 02 25	16.2788	305	9 30 25	17.4642
266	7 07 56	16.3095	306	9 36 36	17.4929
267	7 12 89	16.3401	307	9 42 49	17.5214
268	7 18 24	16.3707	308	9 48 64	17.5499
269	7 23 61	16.4012	309	9 54 81	17.5784
270	7 29 00	16.4317	310	9 61 00	17.6068
271	7 34 41	16.4621	311	9 67 21	17.6352
272	7 39 84	16.4924	312	9 73 44	17.6635
273	7 45 29	16.5227	313	9 79 69	17.6918
274	7 50 76	16.5529	314	9 85 96	17.7200
275	7 56 25	16.5831	315	9 92 25	17.7482
276	7 61 76	16.6132	316	9 98 56	17.7764
277	7 67 29	16.6433	317	10 04 89	17.8045
278	7 72 84	16.6733	318	10 11 24	17.8326
279	7 78 41	16.7033	319	10 17 61	17.8606
280	7 84 00	16.7332	320	10 24 00	17.8885

* From Sorenson. Statistics for students of psychology and education.

TABLE A.—SQUARES AND SQUARE ROOTS OF NUMBERS FROM 1 TO 1,000.*—(*Continued*)

Number	Square	Square root	Number	Square	Square root
321	10 30 41	17.9165	361	13 03 21	19.0000
322	10 36 84	17.9444	362	13 10 44	19.0263
323	10 43 29	17.9722	363	13 17 69	19.0526
324	10 49 76	18.0000	364	13 24 96	19.0788
325	10 56 25	18.0278	365	13 32 25	19.1050
326	10 62 76	18.0555	366	13 39 56	19.1311
327	10 69 29	18.0831	367	13 46 89	19.1572
328	10 75 84	18.1108	368	13 54 24	19.1833
329	10 82 41	18.1384	369	13 61 61	19.2094
330	10 89 00	18.1659	370	13 69 00	19.2354
331	10 95 61	18.1934	371	13 76 41	19.2614
332	11 02 24	18.2209	372	13 83 84	19.2873
333	11 08 89	18.2483	373	13 91 29	19.3132
334	11 15 56	18.2757	374	13 98 76	19.3391
335	11 22 25	18.3030	375	14 06 25	19.3649
336	11 28 96	18.3303	376	14 13 76	19.3907
337	11 35 69	18.3576	377	14 21 29	19.4165
338	11 42 44	18.3848	378	14 28 84	19.4422
339	11 49 21	18.4120	379	14 36 41	19.4679
340	11 56 00	18.4391	380	14 44 00	19.4936
341	11 62 81	18.4662	381	14 51 61	19.5192
342	11 69 64	18.4932	382	14 59 24	19.5448
343	11 76 49	18.5203	383	14 66 89	19.5704
344	11 83 36	18.5472	384	14 74 56	19.5959
345	11 90 25	18.5742	385	14 82 25	19.6214
346	11 97 16	18.6011	386	14 89 96	19.6469
347	12 04 09	18.6279	387	14 97 69	19.6723
348	12 11 04	18.6548	388	15 05 44	19.6977
349	12 18 01	18.6815	389	15 13 21	19.7231
350	12 25 00	18.7083	390	15 21 00	19.7484
351	12 32 01	18.7350	391	15 28 81	19.7737
352	12 39 04	18.7617	392	15 36 64	19.7990
353	12 46 09	18.7883	393	15 44 49	19.8242
354	12 53 16	18.8149	394	15 52 36	19.8494
355	12 60 25	18.8414	395	15 60 25	19.8746
356	12 67 36	18.8680	396	15 68 16	19.8997
357	12 74 49	18.8944	397	15 76 09	19.9249
358	12 81 64	18.9209	398	15 84 04	19.9499
359	12 88 81	18.9473	399	15 92 01	19.9750
360	12 96 00	18.9737	400	16 00 00	20.0000

* From Sorenson. Statistics for students of psychology and education.

TABLE A.—SQUARES AND SQUARE ROOTS OF NUMBERS FROM 1 TO 1,000.*—(*Continued*)

Number	Square	Square root	Number	Square	Square root
401	16 08 01	20.0250	441	19 44 81	21.0000
402	16 16 04	20.0499	442	19 53 64	21.0238
403	16 24 09	20.0749	443	19 62 49	21.0476
404	16 32 16	20.0998	444	19 71 36	21.0713
405	16 40 25	20.1246	445	19 80 25	21.0950
406	16 48 36	20.1494	446	19 89 16	21.1187
407	16 56 49	20.1742	447	19 98 09	21.1424
408	16 64 64	20.1990	448	20 07 04	21.1660
409	16 72 81	20.2237	449	20 16 01	21.1896
410	16 81 00	20.2485	450	20 25 00	21.2132
411	16 89 21	20.2731	451	20 34 01	21.2368
412	16 97 44	20.2978	452	20 43 04	21.2603
413	17 05 69	20.3224	453	20 52 09	21.2838
414	17 13 96	20.3470	454	20 61 16	21.3073
415	17 22 25	20.3715	455	20 70 25	21.3307
416	17 30 56	20.3961	456	20 79 36	21.3542
417	17 38 89	20.4206	457	20 88 49	21.3776
418	17 47 24	20.4450	458	20 97 64	21.4009
419	17 55 61	20.4695	459	21 06 81	21.4243
420	17 64 00	20.4939	460	21 16 00	21.4476
421	17 72 41	20.5183	461	21 25 21	21.4709
422	17 80 84	20.5426	462	21 34 44	21.4942
423	17 89 29	20.5670	463	21 43 69	21.5174
424	17 97 76	20.5913	464	21 52 96	21.5407
425	18 06 25	20.6155	465	21 62 25	21.5639
426	18 14 76	20.6398	466	21 71 56	21.5870
427	18 23 29	20.6640	467	21 80 89	21.6102
428	18 31 84	20.6882	468	21 90 24	21.6333
429	18 40 41	20.7123	469	21 99 61	21.6564
430	18 49 00	20.7364	470	22 09 00	21.6795
431	18 57 61	20.7605	471	22 18 41	21.7025
432	18 66 24	20.7846	472	22 27 84	21.7256
433	18 74 89	20.8087	473	22 37 29	21.7486
434	18 83 56	20.8327	474	22 46 76	21.7715
435	18 92 25	20.8567	475	22 56 25	21.7945
436	19 00 96	20.8806	476	22 65 76	21.8174
437	19 09 69	20.9045	477	22 75 29	21.8403
438	19 18 44	20.9284	478	22 84 84	21.8632
439	19 27 21	20.9523	479	22 94 41	21.8861
440	19 36 00	20.9762	480	23 04 00	21.9089

* From Sorenson. Statistics for students of psychology and education.

TABLE A.—SQUARES AND SQUARE ROOTS OF NUMBERS FROM 1 TO 1,000.*—(*Continued*)

Number	Square	Square root	Number	Square	Square root
481	23 13 61	21.9317	521	27 14 41	22.8254
482	23 23 24	21.9545	522	27 24 84	22.8473
483	23 32 89	21.9773	523	27 35 29	22.8692
484	23 42 56	22.0000	524	27 45 76	22.8910
485	23 52 25	22.0227	525	27 56 25	22.9129
486	23 61 96	22.0454	526	27 66 76	22.9347
487	23 71 69	22.0681	527	27 77 29	22.9565
488	23 81 44	22.0907	528	27 87 84	22.9783
489	23 91 21	22.1133	529	27 98 41	23.0000
490	24 01 00	22.1359	530	28 09 00	23.0217
491	24 10 81	22.1585	531	28 19 61	23.0434
492	24 20 64	22.1811	532	28 30 24	23.0651
493	24 30 49	22.2036	533	28 40 89	23.0868
494	24 40 36	22.2261	534	28 51 56	23.1084
495	24 50 25	22.2486	535	28 62 25	23.1301
496	24 60 16	22.2711	536	28 72 96	23.1517
497	24 70 09	22.2935	537	28 83 69	23.1733
498	24 80 04	22.3159	538	28 94 44	23.1948
499	24 90 01	22.3383	539	29 05 21	23.2164
500	25 00 00	22.3607	540	29 16 00	23.2379
501	25 10 01	22.3830	541	29 26 81	23.2594
502	25 20 04	22.4054	542	29 37 64	23.2809
503	25 30 09	22.4277	543	29 48 49	23.3024
504	25 40 16	22.4499	544	29 59 36	23.3238
505	25 50 25	22.4722	545	29 70 25	23.3452
506	25 60 36	22.4944	546	29 81 16	23.3666
507	25 70 49	22.5167	547	29 92 09	23.3880
508	25 80 64	22.5389	548	30 03 04	23.4094
509	25 90 81	22.5610	549	30 14 01	23.4307
510	26 01 00	22.5832	550	30 25 00	23.4521
511	26 11 21	22.6053	551	30 36 01	23.4734
512	26 21 44	22.6274	552	30 47 04	23.4947
513	26 31 69	22.6495	553	30 58 09	23.5160
514	26 41 96	22.6716	554	30 69 16	23.5372
515	26 52 25	22.6936	555	30 80 25	23.5584
516	26 62 56	22.7156	556	30 91 36	23.5797
517	26 72 89	22.7376	557	31 02 49	23.6008
518	26 83 24	22.7596	558	31 13 64	23.6220
519	26 93 61	22.7816	559	31 24 81	23.6432
520	27 04 00	22.8035	560	31 36 00	23.6643

* From Sorenson. Statistics for students of psychology and education.

TABLE A.—SQUARES AND SQUARE ROOTS OF NUMBERS FROM 1 TO 1,000.*—(*Continued*)

Number	Square	Square root	Number	Square	Square root
561	31 47 21	23.6854	601	36 12 01	24.5153
562	31 58 44	23.7065	602	36 24 04	24.5357
563	31 69 69	23.7276	603	36 36 09	24.5561
564	31 80 96	23.7487	604	36 48 16	24.5764
565	31 92 25	23.7697	605	36 60 25	24.5967
566	32 03 56	23.7908	606	36 72 36	24.6171
567	32 14 89	23.8118	607	36 84 49	24.6374
568	32 26 24	23.8328	608	36 96 64	24.6577
569	32 37 61	23.8537	609	37 08 81	24.6779
570	32 49 00	23.8747	610	37 21 00	24.6982
571	32 60 41	23.8956	611	37 33 21	24.7184
572	32 71 84	23.9165	612	37 45 44	24.7385
573	32 83 29	23.9374	613	37 57 69	24.7588
574	32 94 76	23.9583	614	37 69 96	24.7790
575	33 06 25	23.9792	615	37 82 25	24.7992
576	33 17 76	24.0000	616	37 94 56	24.8193
577	33 29 29	24.0208	617	38 06 89	24.8395
578	33 40 84	24.0416	618	38 19 24	24.8596
579	33 52 41	24.0624	619	38 31 61	24.8797
580	33 64 00	24.0832	620	38 44 00	24.8998
581	33 75 61	24.1039	621	38 56 41	24.9199
582	33 87 24	24.1247	622	38 68 84	24.9399
583	33 98 89	24.1454	623	38 81 29	24.9600
584	34 10 56	24.1661	624	38 93 76	24.9800
585	34 22 25	24.1868	625	39 06 25	25.0000
586	34 33 96	24.2074	626	39 18 76	25.0200
587	34 45 69	24.2281	627	39 31 29	25.0400
588	34 57 44	24.2487	628	39 43 84	25.0599
589	34 69 21	24.2693	629	39 56 41	25.0799
590	34 81 00	24.2899	630	39 69 00	25.0998
591	34 92 81	24.3105	631	39 81 61	25.1197
592	35 04 64	24.3311	632	39 94 24	25.1396
593	35 16 49	24.3516	633	40 06 89	25.1595
594	35 28 36	24.3721	634	40 19 56	25.1794
595	35 40 25	24.3926	635	40 32 25	25.1992
596	35 52 16	24.4131	636	40 44 96	25.2190
597	35 64 09	24.4336	637	40 57 69	25.2389
598	35 76 04	24.4540	638	40 70 44	25.2587
599	35 88 01	24.4745	639	40 83 21	25.2784
600	36 00 00	24.4949	640	40 96 00	25.2982

* From Sorenson. Statistics for students of psychology and education.

TABLE A.—SQUARES AND SQUARE ROOTS OF NUMBERS FROM 1 TO 1,000.*—(*Continued*)

Number	Square	Square root	Number	Square	Square root
641	41 08 81	25.3180	681	46 37 61	26.0960
642	41 21 64	25.3377	682	46 51 24	26.1151
643	41 34 49	25.3574	683	46 64 89	26.1343
644	41 47 36	25.3772	684	46 78 56	26.1534
645	41 60 25	25.3969	685	46 92 25	26.1725
646	41 73 16	25.4165	686	47 05 96	26.1916
647	41 86 09	25.4362	687	47 19 69	26.2107
648	41 99 04	25.4558	688	47 33 44	26.2298
649	42 12 01	25.4755	689	47 47 21	26.2488
650	42 25 00	25.4951	690	47 61 00	26.2679
651	42 38 01	25.5147	691	47 74 81	26.2869
652	42 51 04	25.5343	692	47 88 64	26.3059
653	42 64 09	25.5539	693	48 02 49	26.3249
654	42 77 16	25.5734	694	48 16 36	26.3439
655	42 90 25	25.5930	695	48 30 25	26.3629
656	43 03 36	25.6125	696	48 44 16	26.3818
657	43 16 49	25.6320	697	48 58 09	26.4008
658	43 29 64	25.6515	698	48 72 04	26.4197
659	43 42 81	25.6710	699	48 86 01	26.4386
660	43 56 00	25.6905	700	49 00 00	26.4575
661	43 69 21	25.7099	701	49 14 01	26.4764
662	43 82 44	25.7294	702	49 28 04	26.4953
663	43 95 69	25.7488	703	49 42 09	26.5141
664	44 08 96	25.7682	704	49 56 16	26.5330
665	44 22 25	25.7876	705	49 70 25	26.5518
666	44 35 56	25.8070	706	49 84 36	26.5707
667	44 48 89	25.8263	707	49 98 49	26.5895
668	44 62 24	25.8457	708	50 12 64	26.6083
669	44 75 61	25.8650	709	50 26 81	26.6271
670	44 89 00	25.8844	710	50 41 00	26.6458
671	45 02 41	25.9037	711	50 55 21	26.6646
672	45 15 84	25.9230	712	50 69 44	26.6833
673	45 29 29	25.9422	713	50 83 69	26.7021
674	45 42 76	25.9615	714	50 97 96	26.7208
675	45 56 25	25.9808	715	51 12 25	26.7395
676	45 69 76	26.0000	716	51 26 56	26.7582
677	45 83 29	26.0192	717	51 40 89	26.7769
678	45 96 84	26.0384	718	51 55 24	26.7955
679	46 10 41	26.0576	719	51 69 61	26.8142
680	46 24 00	26.0768	720	51 84 00	26.8328

* From Sorenson. Statistics for students of psychology and education.

TABLE A.—SQUARES AND SQUARE ROOTS OF NUMBERS FROM 1 TO 1,000.*—(*Continued*)

Number	Square	Square root	Number	Square	Square root
721	51 98 41	26.8514	761	57 91 21	27.5862
722	52 12 84	26.8701	762	58 06 44	27.6043
723	52 27 29	26.8887	763	58 21 69	27.6225
724	52 41 76	26.9072	764	58 36 96	27.6405
725	52 56 25	26.9258	765	58 52 25	27.6586
726	52 70 76	26.9444	766	58 67 56	27.6767
727	52 85 29	26.9629	767	58 82 89	27.6948
728	52 99 84	26.9815	768	58 98 24	27.7128
729	53 14 41	27.0000	769	59 13 61	27.7308
730	53 29 00	27.0185	770	59 29 00	27.7489
731	53 43 61	27.0370	771	59 44 41	27.7669
732	53 58 24	27.0555	772	59 59 84	27.7849
733	53 72 89	27.0740	773	59 75 29	27.8029
734	53 87 56	27.0924	774	59 90 76	27.8209
735	54 02 25	27.1109	775	60 06 25	27.8388
736	54 16 96	27.1293	776	60 21 76	27.8568
737	54 31 69	27.1477	777	60 37 29	27.8747
738	54 46 44	27.1662	778	60 52 84	27.8927
739	54 61 27	27.1846	779	60 68 41	27.9106
740	54 76 00	27.2029	780	60 84 00	27.9285
741	54 90 81	27.2213	781	60 99 61	27.9464
742	55 05 64	27.2397	782	61 15 24	27.9643
743	55 20 49	27.2580	783	61 30 89	27.9821
744	55 35 36	27.2764	784	61 46 56	28.0000
745	55 50 25	27.2947	785	61 62 25	28.0179
746	55 65 16	27.3130	786	61 77 96	28.0357
747	55 80 09	27.3313	787	61 93 69	28.0535
748	55 95 04	27.3496	788	62 09 44	28.0713
749	56 10 01	27.3679	789	62 25 21	28.0891
750	56 25 00	27.3861	790	62 41 00	28.1069
751	56 40 01	27.4044	791	62 56 81	28.1247
752	56 55 04	27.4226	792	62 72 64	28.1425
753	56 70 09	27.4408	793	62 88 49	28.1603
754	56 85 16	27.4591	794	63 04 36	28.1780
755	57 00 25	27.4773	795	63 20 25	28.1957
756	57 15 36	27.4955	796	63 36 16	28.2135
757	57 30 49	27.5136	797	63 52 09	28.2312
758	57 45 64	27.5318	798	63 68 04	28.2489
759	57 60 81	27.5500	799	63 84 01	28.2666
760	57 76 00	27.5681	800	64 00 00	28.2843

* From Sorenson. Statistics for students of psychology and education.

TABLE A.—SQUARES AND SQUARE ROOTS OF NUMBERS FROM 1 TO 1,000.*—(*Continued*)

Number	Square	Square root	Number	Square	Square root
801	64 16 01	28.3019	841	70 72 81	29.0000
802	64 32 04	28.3196	842	70 89 64	29.0172
803	64 48 09	28.3373	843	71 06 49	29.0345
804	64 64 16	28.3049	844	71 23 36	29.0517
805	64 80 25	28.3725	845	71 40 25	29.0689
806	64 96 36	28.3901	846	71 57 16	29.0861
807	65 12 49	28.4077	847	71 74 09	29.1033
808	65 28 64	28.4253	848	71 91 04	29.1204
809	65 44 81	28.4429	849	72 08 01	29.1376
810	65 61 00	28.4605	850	72 25 00	29.1548
811	65 77 21	28.4781	851	72 42 01	29.1719
812	65 93 44	28.4956	852	72 59 04	29.1890
813	66 09 69	28.5132	853	72 76 09	29.2062
814	66 25 96	28.5307	854	72 93 16	29.2233
815	66 42 25	28.5482	855	73 10 25	29.2404
816	66 58 56	28.5657	856	73 27 36	29.2575
817	66 74 89	28.5832	857	73 44 49	29.2746
818	66 91 24	28.6007	858	73 61 64	29.2916
819	67 07 61	28.6082	859	73 78 81	29.3087
820	67 24 00	28.6356	860	73 96 00	29.3258
821	67 40 41	28.6531	861	74 13 21	29.3428
822	67 56 84	28.6705	862	74 30 44	29.3598
823	67 73 29	28.6880	863	74 47 69	29.3769
824	67 89 76	28.7054	864	74 64 96	29.3939
825	68 06 25	28.7228	865	74 82 25	29.4109
826	68 22 76	28.7402	866	74 99 56	29.4279
827	68 39 29	28.7576	867	75 16 89	29.4449
828	68 55 84	28.7750	868	75 34 24	29.4618
829	68 72 41	28.7924	869	75 51 61	29.4788
830	68 89 00	28.8097	870	75 69 00	29.4958
831	69 05 61	28.8271	871	75 86 41	29.5127
832	69 22 24	28.8444	872	76 03 84	29.5296
833	69 38 89	28.8617	873	76 21 29	29.5466
834	69 55 56	28.8791	874	76 38 76	29.5635
835	69 72 25	28.8964	875	76 56 25	29.5804
836	69 88 96	28.9137	876	76 73 76	29.5973
837	70 05 69	28.9310	877	76 91 29	29.6142
838	70 22 44	28.9482	878	77 08 84	29.6311
839	70 39 21	28.9655	879	77 26 41	29.6479
840	70 56 00	28.9828	880	77 44 00	29.6648

* From Sorenson. Statistics for students of psychology and education.

Table A.—Squares and Square Roots of Numbers from 1 to 1,000.*—(*Continued*)

Number	Square	Square root	Number	Square	Square root
881	77 61 61	29.6816	921	84 82 41	30.3480
882	77 79 24	29.6985	922	85 00 84	30.3645
883	77 96 89	29.7153	923	85 19 29	30.3809
884	78 14 56	29.7321	924	85 37 76	30.3974
885	78 32 25	29.7489	925	85 56 25	30.4138
886	78 49 96	29.7658	926	85 74 76	30.4302
887	78 67 69	29.7825	927	85 93 29	30.4467
888	78 85 44	29.7993	928	86 11 84	30.4631
889	79 03 21	29.8161	929	86 30 41	30.4795
890	79 21 00	29.8329	930	86 49 00	30.4959
891	79 38 81	29.8496	931	86 67 61	30.5123
892	79 56 64	29.8664	932	86 86 24	30.5287
893	79 74 49	29.8831	933	87 04 89	30.5450
894	79 92 36	29.8998	934	87 23 56	30.5614
895	80 10 25	29.9166	935	87 42 25	30.5778
896	80 28 16	29.9333	936	87 60 96	30.5941
897	80 46 09	29.9500	937	87 79 69	30.6105
898	80 64 04	29.9666	938	87 98 44	30.6268
899	80 82 01	29.9833	939	88 17 21	30.6431
900	81 00 00	30.0000	940	88 36 00	30.6594
901	81 18 01	30.0167	941	88 54 81	30.6757
902	81 36 04	30.0333	942	88 73 64	30.6920
903	81 54 09	30.0500	943	88 92 49	30.7083
904	81 72 16	30.0666	944	89 11 36	30.7246
905	81 90 25	30.0832	945	89 30 25	30.7409
906	82 08 36	30.0998	946	89 49 16	30.7571
907	82 26 49	30.1164	947	89 68 09	30.7734
908	82 44 64	30.1330	948	89 87 04	30.7896
909	82 62 81	30.1496	949	90 06 01	30.8058
910	82 81 00	30.1662	950	90 25 00	30.8221
911	82 99 21	30.1828	951	90 44 01	30.8383
912	83 17 44	30.1993	952	90 63 04	30.8545
913	83 35 69	30.2159	953	90 82 09	30.8707
914	83 53 96	30.2324	954	91 01 16	30.8869
915	83 72 25	30.2490	955	91 20 25	30.9031
916	83 90 56	30.2655	956	91 39 36	30.9192
917	84 08 89	30.2820	957	91 58 49	30.9354
918	84 27 24	30.2985	958	91 77 64	30.9516
919	84 45 61	30.3150	959	91 96 81	30.9677
920	84 64 00	30.3315	960	92 16 00	30.9839

* From Sorenson. Statistics for students of psychology and education.

TABLE A.—SQUARES AND SQUARE ROOTS OF NUMBERS FROM 1 TO 1,000.*—(*Continued*)

Number	Square	Square root	Number	Square	Square root
961	92 35 21	31.0000	981	96 23 61	31.3209
962	92 54 44	31.0161	982	96 43 24	31.3369
963	92 73 69	31.0322	983	96 62 89	31.3528
964	92 92 96	31.0483	984	96 82 56	31.3688
965	93 12 25	31.0644	985	97 02 25	31.3847
966	93 31 56	31.0805	986	97 21 96	31.4006
967	93 50 89	31.0966	987	97 41 69	31.4166
968	93 70 24	31.1127	988	97 61 44	31.4325
969	93 89 61	31.1288	989	97 81 21	31.4484
970	94 09 00	31.1448	990	98 01 00	31.4643
971	94 28 41	31.1609	991	98 20 81	31.4802
972	94 47 84	31.1769	992	98 40 64	31.4960
973	94 67 29	31.1929	993	98 60 49	31.5119
974	94 86 76	31.2090	994	98 80 36	31.5278
975	95 06 25	31.2250	995	99 00 25	31.5436
976	95 25 76	31.2410	996	99 20 16	31.5595
977	95 45 29	31.2570	997	99 40 09	31.5753
978	95 64 84	31.2730	998	99 60 04	31.5911
979	95 84 41	31.2890	999	99 80 01	31.6070
980	96 04 00	31.3050	1000	100 00 00	31.6228

* From Sorenson, H. Statistics for students of psychology and education. New York: McGraw-Hill, 1936.

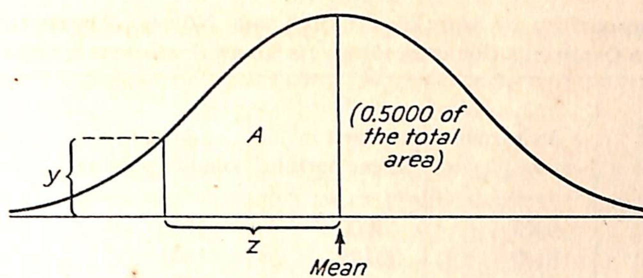
The Use of Tables B and C

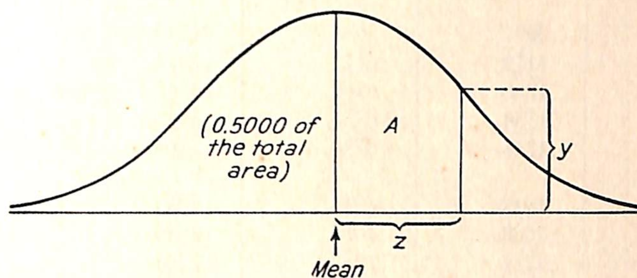
Tables B and C assume a normal distribution whose standard deviation is equal to 1.00 and whose total area (or N) also equals 1.00. Under these conditions, there are fixed mathematical relationships between values on the base line (as measured in sigma units) and areas under the curve (A, B, and C) and also ordinate values (y).

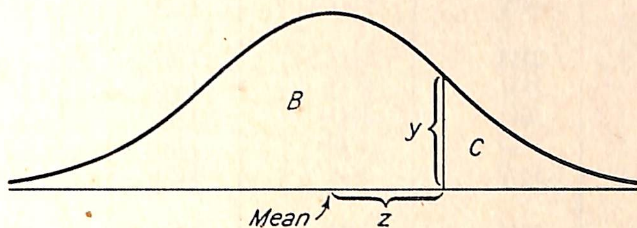
The use of Tables B and C is fully explained in Ch. 7. Figs. A.1, A.2, B.1, and B.2 may help to relate the symbols to the normal curve.

Table B is best used when we know a z and want to find a corresponding A, B, or C area, or the ordinate y. Table C is best used when we know any one of the areas A, B, or C and want to find the corresponding z or y. In case any one of these areas is known, it can be readily used to find a corresponding area by the use of the following relationships.

$$A = B - .50$$
$$A = .50 - C \qquad (A + C = .50)$$
$$B = A + .50$$
$$B = 1.00 - C \qquad (B + C = 1.00)$$
$$C = .50 - A$$
$$C = 1.00 - B$$


(0.5000 of
the total
area)
A
y
z
Mean


(0.5000 of
the total
area)
A
y
z
Mean


B
y
C
Mean
z

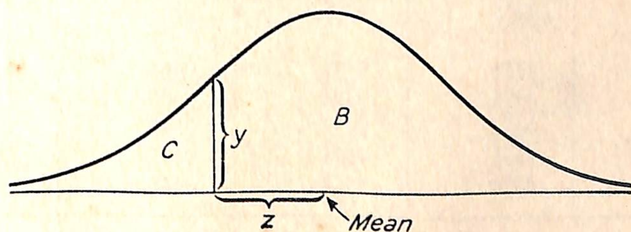

B
C
y
z
Mean

TABLE B.—PROPORTIONS OF THE AREA UNDER THE NORMAL DISTRIBUTION CURVE AND ORDINATES CORRESPONDING TO GIVEN STANDARD SCORES

z Standard score (x/σ)	A Area from mean to x/σ	B Area in larger portion	C Area in smaller portion	y Ordinate at x/σ
0.00	.0000	.5000	.5000	.3989
0.05	.0199	.5199	.4801	.3984
0.10	.0398	.5398	.4602	.3970
0.15	.0596	.5596	.4404	.3945
0.20	.0793	.5793	.4207	.3910
0.25	.0987	.5987	.4013	.3867
0.30	.1179	.6179	.3821	.3814
0.35	.1368	.6368	.3632	.3752
0.40	.1554	.6554	.3446	.3683
0.45	.1736	.6736	.3264	.3605
0.50	.1915	.6915	.3085	.3521
0.55	.2088	.7088	.2912	.3429
0.60	.2257	.7257	.2743	.3332
0.65	.2422	.7422	.2578	.3230
0.70	.2580	.7580	.2420	.3123
0.75	.2734	.7734	.2266	.3011
0.80	.2881	.7881	.2119	.2897
0.85	.3023	.8023	.1977	.2780
0.90	.3159	.8159	.1841	.2661
0.95	.3289	.8289	.1711	.2541
1.00	.3413	.8413	.1587	.2420
1.05	.3531	.8531	.1469	.2299
1.10	.3643	.8643	.1357	.2179
1.15	.3749	.8749	.1251	.2059
1.20	.3849	.8849	.1151	.1942
1.25	.3944	.8944	.1056	.1826
1.30	.4032	.9032	.0968	.1714
1.35	.4115	.9115	.0885	.1604
1.40	.4192	.9192	.0808	.1497
1.45	.4265	.9265	.0735	.1394
1.50	.4332	.9332	.0668	.1295
1.55	.4394	.9394	.0606	.1200
1.60	.4452	.9452	.0548	.1109
1.65	.4505	.9505	.0495	.1023
1.70	.4554	.9554	.0446	.0940

TABLE B.—PROPORTIONS OF THE AREA UNDER THE NORMAL DISTRIBUTION CURVE AND ORDINATES CORRESPONDING TO GIVEN STANDARD SCORES.—(*Continued*)

z Standard score (x/σ)	A Area from mean to x/σ	B Area in larger portion	C Area in smaller portion	y Ordinate at x/σ
1.75	.4599	.9599	.0401	.0863
1.80	.4641	.9641	.0359	.0790
1.85	.4678	.9678	.0322	.0721
1.90	.4713	.9713	.0287	.0656
1.95	.4744	.9744	.0256	.0596
2.00	.4772	.9772	.0228	.0540
2.05	.4798	.9798	.0202	.0488
2.10	.4821	.9821	.0179	.0440
2.15	.4842	.9842	.0158	.0396
2.20	.4861	.9861	.0139	.0355
2.25	.4878	.9878	.0122	.0317
2.30	.4893	.9893	.0107	.0283
2.35	.4906	.9906	.0094	.0252
2.40	.4918	.9918	.0082	.0224
2.45	.4929	.9929	.0071	.0198
2.50	.4938	.9938	.0062	.0175
2.55	.4946	.9946	.0054	.0154
2.60	.4953	.9953	.0047	.0136
2.65	.4960	.9960	.0040	.0119
2.70	.4965	.9965	.0035	.0104
2.80	.4974	.9974	.0026	.0079
2.90	.4981	.9981	.0019	.0060
3.00	.49865	.99865	.00135	.0044
3.10	.49903	.99903	.00097	.0033
3.20	.49931	.99931	.00069	.0024
3.40	.49966	.99966	.00034	.0012
3.60	.49984	.99984	.00016	.00061
3.80	.499928	.999928	.000072	.00029
4.00	.4999683	.9999683	.0000317	.00013
4.50	.4999966	.9999966	.0000034	.000015
5.00	.49999971	.99999971	.00000029	.0000015
6.00	.499999999	.999999999	.000000001	.000000006

TABLE C.—STANDARD SCORES (OR DEVIATES) AND ORDINATES CORRESPONDING TO DIVISIONS OF THE AREA UNDER THE NORMAL CURVE INTO A LARGER PROPORTION (B) AND A SMALLER PROPORTION (C); ALSO THE VALUE $\sqrt{BC}$

B The larger area	z Standard score	y Ordinate	$\sqrt{BC}$	C The smaller area
.500	.0000	.3989	.5000	.500
.505	.0125	.3989	.5000	.495
.510	.0251	.3988	.4999	.490
.515	.0376	.3987	.4998	.485
.520	.0502	.3984	.4996	.480
.525	.0627	.3982	.4994	.475
.530	.0753	.3978	.4991	.470
.535	.0878	.3974	.4988	.465
.540	.1004	.3969	.4984	.460
.545	.1130	.3964	.4980	.455
.550	.1257	.3958	.4975	.450
.555	.1383	.3951	.4970	.445
.560	.1510	.3944	.4964	.440
.565	.1637	.3936	.4958	.435
.570	.1764	.3928	.4951	.430
.575	.1891	.3919	.4943	.425
.580	.2019	.3909	.4936	.420
.585	.2147	.3899	.4927	.415
.590	.2275	.3887	.4918	.410
.595	.2404	.3876	.4909	.405
.600	.2533	.3863	.4899	.400
.605	.2663	.3850	.4889	.395
.610	.2793	.3837	.4877	.390
.615	.2924	.3822	.4867	.385
.620	.3055	.3808	.4854	.380
.625	.3186	.3792	.4841	.375
.630	.3319	.3776	.4828	.370
.635	.3451	.3759	.4814	.365
.640	.3585	.3741	.4800	.360
.645	.3719	.3723	.4785	.355
.650	.3853	.3704	.4770	.350
.655	.3989	.3684	.4754	.345
.660	.4125	.3664	.4737	.340
.665	.4261	.3643	.4720	.335
.670	.4399	.3621	.4702	.330
.675	.4538	.3599	.4684	.325
.680	.4677	.3576	.4665	.320
.685	.4817	.3552	.4645	.315
.690	.4959	.3528	.4625	.310
.695	.5101	.3503	.4604	.305
.700	.5244	.3477	.4583	.300
705	.5388	.3450	.4560	.295
.710	.5534	.3423	.4538	.290
.715	.5681	.3395	.4514	.285
.720	.5828	.3366	.4490	.280

Table C.—Standard Scores (or Deviates) and Ordinates Corresponding to Divisions of the Area under the Normal Curve into a Larger Proportion (B) and a Smaller Proportion (C); also the Value $\sqrt{BC}$.—(*Continued*)

B The larger area	z Standard score	y Ordinate	$\sqrt{BC}$	C The smaller area
.725	.5978	.3337	.4465	.275
.730	.6128	.3306	.4440	.270
.735	.6280	.3275	.4413	.265
.740	.6433	.3244	.4386	.260
.745	.6588	.3211	.4359	.255
.750	.6745	.3178	.4330	.250
.755	.6903	.3144	.4301	.245
.760	.7063	.3109	.4271	.240
.765	.7225	.3073	.4240	.235
.770	.7388	.3036	.4208	.230
.775	.7554	.2999	.4176	.225
.780	.7722	.2961	.4142	.220
.785	.7892	.2922	.4108	.215
.790	.8064	.2882	.4073	.210
.795	.8239	.2841	.4037	.205
.800	.8416	.2800	.4000	.200
.805	.8596	.2757	.3962	.195
.810	.8779	.2714	.3923	.190
.815	.8965	.2669	.3883	.185
.820	.9154	.2624	.3842	.180
.825	.9346	.2578	.3800	.175
.830	.9542	.2531	.3756	.170
.835	.9741	.2482	.3712	.165
.840	.9945	.2433	.3666	.160
.845	1.0152	.2383	.3619	.155
.850	1.0364	.2332	.3571	.150
.855	1.0581	.2279	.3521	.145
.860	1.0803	.2226	.3470	.140
.865	1.1031	.2171	.3417	.135
.870	1.1264	.2115	.3363	.130
.875	1.1503	.2059	.3307	.125
.880	1.1750	.2000	.3250	.120
.885	1.2004	.1941	.3190	.115
.890	1.2265	.1880	.3129	.110
.895	1.2536	.1818	.3066	.105
.900	1.2816	.1755	.3000	.100
.905	1.3016	.1690	.2932	.095
.910	1.3408	.1624	.2862	.090
.915	1.3722	.1556	.2789	.085
.920	1.4051	.1487	.2713	.080
.925	1.4395	.1416	.2634	.075
.930	1.4757	.1343	.2551	.070
.935	1.5141	.1268	.2465	.065
.940	1.5548	.1191	.2375	.060
.945	1.5982	.1112	.2280	.055

TABLE C.—STANDARD SCORES (OR DEVIATES) AND ORDINATES CORRESPONDING TO DIVISIONS OF THE AREA UNDER THE NORMAL CURVE INTO A LARGER PROPORTION (B) AND A SMALLER PROPORTION (C); ALSO THE VALUE $\sqrt{BC}$.—(*Continued*)

B The larger area	z Standard score	y Ordinate	$\sqrt{BC}$	C The smaller area
.950	1.6449	.1031	.2179	.050
.955	1.6954	.0948	.2073	.045
.960	1.7507	.0862	.1960	.040
.965	1.8119	.0773	.1838	.035
.970	1.8808	.0680	.1706	.030
.975	1.9600	.0584	.1561	.025
.980	2.0537	.0484	.1400	.020
.985	2.1701	.0379	.2126	.015
.990	2.3263	.0267	.0995	.010
.995	2.5758	.0145	.0705	.005
.996	2.6521	.0118	.0631	.004
.997	2.7478	.0091	.0547	.003
998	2.8782	.0063	.0447	.002
999	3.0902	.0034	.0316	.001
.9995	3.2905	.0018	.0224	.0005

TABLE D.—COEFFICIENTS OF CORRELATION AND *t* RATIOS SIGNIFICANT AT THE 5 PER CENT LEVEL (ROMAN TYPE) AND AT THE 1 PER CENT LEVEL (BOLD-FACED TYPE) FOR VARYING DEGREES OF FREEDOM*

Degrees of freedom	Number of variables									*t*
	2	3	4	5	6	7	9	13	25	
1	.997	.999	.999	.999	1.000	1.000	1.000	1.000	1.000	12.706
	1.000	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**63.657**
2	.950	.975	.983	.987	.990	.992	.994	.996	.998	4.303
	.990	**.995**	**.997**	**.998**	**.998**	**.998**	**.999**	**.999**	**1.000**	**9.925**
3	.878	.930	.950	.961	.968	.973	.979	.986	.993	3.182
	.959	**.976**	**.983**	**.987**	**.990**	**.991**	**.993**	**.995**	**.998**	**5.841**
4	.811	.881	.912	.930	.942	.950	.961	.973	.986	2.776
	.917	**.949**	**.962**	**.970**	**.975**	**.979**	**.984**	**.989**	**.994**	**4.604**
5	.754	.836	.874	.898	.914	.925	.941	.958	.978	2.571
	.874	**.917**	**.937**	**.949**	**.957**	**.963**	**.971**	**.980**	**.989**	**4.032**
6	.707	.795	.839	.867	.886	.900	.920	.943	.969	2.447
	.834	**.886**	**.911**	**.927**	**.938**	**.946**	**.957**	**.969**	**.983**	**3.707**
7	.666	.758	.807	.838	.860	.876	.900	.927	.960	2.365
	.798	**.855**	**.885**	**.904**	**.918**	**.928**	**.942**	**.958**	**.977**	**3.499**
8	.632	.726	.777	.811	.835	.854	.880	.912	.950	2.306
	.765	**.827**	**.860**	**.882**	**.898**	**.909**	**.926**	**.946**	**.970**	**3.355**
9	.602	.697	.750	.786	.812	.832	.861	.897	.941	2.262
	.735	**.800**	**.836**	**.861**	**.878**	**.891**	**.911**	**.934**	**.963**	**3.250**
10	.576	.671	.726	.763	.790	.812	.843	.882	.932	2.228
	.708	**.776**	**.814**	**.840**	**.859**	**.874**	**.895**	**.922**	**.955**	**3.169**
11	.553	.648	.703	.741	.770	.792	.826	.868	.922	2.201
	.684	**.753**	**.793**	**.821**	**.841**	**.857**	**.880**	**.910**	**.948**	**3.106**
12	.532	.627	.683	.722	.751	.774	.809	.854	.913	2.179
	.661	**.732**	**.773**	**.802**	**.824**	**.841**	**.866**	**.898**	**.940**	**3.055**
13	.514	.608	.664	.703	.733	.757	.794	.840	.904	2.160
	.641	**.712**	**.755**	**.785**	**.807**	**.825**	**.852**	**.886**	**.932**	**3.012**
14	.497	.590	.646	.686	.717	.741	.779	.828	.895	2.145
	.623	**.694**	**.737**	**.768**	**.792**	**.810**	**.838**	**.875**	**.924**	**2.977**
15	.482	.574	.630	.670	.701	.726	.765	.815	.886	2.131
	.606	**.677**	**.721**	**.752**	**.776**	**.796**	**.825**	**.864**	**.917**	**2.947**
16	.468	.559	.615	.655	.686	.712	.751	.803	.878	2.120
	.590	**.662**	**.706**	**.738**	**.762**	**.782**	**.813**	**.853**	**.909**	**2.921**
17	.456	.545	.601	.641	.673	.698	.738	.792	.869	2.110
	.575	**.647**	**.691**	**.724**	**.749**	**.769**	**.800**	**.842**	**.902**	**2.898**
18	.444	.532	.587	.628	.660	.686	.726	.781	.861	2.101
	.561	**.633**	**.678**	**.710**	**.736**	**.756**	**.789**	**.832**	**.894**	**2.878**
19	.433	.520	.575	.615	.647	.674	.714	.770	.853	2.093
	.549	**.620**	**.665**	**.698**	**.723**	**.744**	**.778**	**.822**	**.887**	**2.861**
20	.423	.509	.563	.604	.636	.662	.703	.760	.845	2.086
	.537	**.608**	**.652**	**.685**	**.712**	**.733**	**.767**	**.812**	**.880**	**2.845**
21	.413	.498	.552	.592	.624	.651	.693	.750	.837	2.080
	.526	**.596**	**.641**	**.674**	**.700**	**.722**	**.756**	**.803**	**.873**	**2.831**
22	.404	.488	.542	.582	.614	.640	.682	.740	.830	2.074
	.515	**.585**	**.630**	**.663**	**.690**	**.712**	**.746**	**.794**	**.866**	**2.819**
23	.396	.479	.532	.572	.604	.630	.673	.731	.823	2.069
	.505	**.574**	**.619**	**.652**	**.679**	**.701**	**.736**	**.785**	**.859**	**2.807**

*Adopted from Wallace, H. A., and Snedecor, G. W. Correlation and machine calculation, 1931, by courtesy of the authors.

TABLE D.—COEFFICIENTS OF CORRELATION AND *t* RATIOS SIGNIFICANT AT THE 5 PER CENT LEVEL (ROMAN TYPE) AND AT THE 1 PER CENT LEVEL (BOLD-FACED TYPE) FOR VARYING DEGREES OF FREEDOM.*—(*Continued*)

Degrees of freedom	Number of variables									*t*
	2	3	4	5	6	7	9	13	25	
24	.388	.470	.523	.562	.594	.621	.663	.722	.815	2.064
	.496	**.565**	**.609**	**.642**	**.669**	**.692**	**.727**	**.776**	**.852**	**2.797**
25	.381	.462	.514	.553	.585	.612	.654	.714	.808	2.060
	.487	**.555**	**.600**	**.633**	**.660**	**.682**	**.718**	**.768**	**.846**	**2.787**
26	.374	.454	.506	.545	.576	.603	.645	.706	.802	2.056
	.478	**.546**	**.590**	**.624**	**.651**	**.673**	**.709**	**.760**	**.839**	**2.779**
27	.367	.446	.498	.536	.568	.594	.637	.698	.795	2.052
	.470	**.538**	**.582**	**.615**	**.642**	**.664**	**.701**	**.752**	**.833**	**2.771**
28	.361	.439	.490	.529	.560	.586	.629	.690	.788	2.048
	.463	**.530**	**.573**	**.606**	**.634**	**.656**	**.692**	**.744**	**.827**	**2.763**
29	.355	.432	.482	.521	.552	.579	.621	.682	.782	2.045
	.456	**.522**	**.565**	**.598**	**.625**	**.648**	**.685**	**.737**	**.821**	**2.756**
30	.349	.426	.476	.514	.545	.571	.614	.675	.776	2.042
	.449	**.514**	**.558**	**.591**	**.618**	**.640**	**.677**	**.729**	**.815**	**2.750**
35	.325	.397	.445	.482	.512	.538	.580	.642	.746	2.030
	.418	**.481**	**.523**	**.556**	**.582**	**.605**	**.642**	**.696**	**.786**	**2.724**
40	.304	.373	.419	.455	.484	.509	.551	.613	.720	2.021
	.393	**.454**	**.494**	**.526**	**.552**	**.575**	**.612**	**.667**	**.761**	**2.704**
45	.288	.353	.397	.432	.460	.485	.526	.587	.696	2.014
	.372	**.430**	**.470**	**.501**	**.527**	**.549**	**.586**	**.640**	**.737**	**2.690**
50	.273	.336	.379	.412	.440	.464	.504	.565	.674	2.008
	.354	**.410**	**.449**	**.479**	**.504**	**.526**	**.562**	**.617**	**.715**	**2.678**
60	.250	.308	.348	.380	.406	.429	.467	.526	.636	2.000
	.325	**.377**	**.414**	**.442**	**.466**	**.488**	**.523**	**.577**	**.677**	**2.660**
70	.233	.286	.324	.354	.379	.401	.438	.495	.604	1.994
	.302	**.351**	**.386**	**.413**	**.436**	**.456**	**.491**	**.544**	**.644**	**2.648**
80	.217	.269	.304	.332	.356	.377	.413	.469	.576	1.990
	.283	**.330**	**.362**	**.389**	**.411**	**.431**	**.464**	**.516**	**.615**	**2.638**
90	.205	.254	.288	.315	.338	.358	.392	.446	.552	1.987
	.267	**.312**	**.343**	**.368**	**.390**	**.409**	**.441**	**.492**	**.590**	**2.632**
100	.195	.241	.274	.300	.322	.341	.374	.426	.530	1.984
	.254	**.297**	**.327**	**.351**	**.372**	**.390**	**.421**	**.470**	**.568**	**2.626**
125	.174	.216	.246	.269	.290	.307	.338	.387	.485	1.979
	.228	**.266**	**.294**	**.316**	**.335**	**.352**	**.381**	**.428**	**.521**	**2.616**
150	.159	.198	.225	.247	.266	.282	.310	.356	.450	1.976
	.208	**.244**	**.270**	**.290**	**.308**	**.324**	**.351**	**.395**	**.484**	**2.609**
200	.138	.172	.196	.215	.231	.246	.271	.312	.398	1.972
	.181	**.212**	**.234**	**.253**	**.269**	**.283**	**.307**	**.347**	**.430**	**2.601**
300	.113	.141	.160	.176	.190	.202	.223	.258	.332	1.968
	.148	**.174**	**.192**	**.208**	**.221**	**.233**	**.253**	**.287**	**.359**	**2.592**
400	.098	.122	.139	.153	.165	.176	.194	.225	.291	1.966
	.128	**.151**	**.167**	**.180**	**.192**	**.202**	**.220**	**.250**	**.315**	**2.588**
500	.088	.109	.124	.137	.148	.157	.174	.202	.262	1.965
	.115	**.135**	**.150**	**.162**	**.172**	**.182**	**.198**	**.225**	**.284**	**2.586**
1000	.062	.077	.088	.097	.105	.112	.124	.144	.188	1.962
	.081	**.096**	**.106**	**.115**	**.122**	**.129**	**.141**	**.160**	**.204**	**2.581**
∞										1.960
										2.576

TABLE E.—TABLE OF CHI SQUARE*

n	P = .99	.98	.95	.90	.80	.70	.50	.30	.20	.10	.05	.02	.01
1	.000157	.000628	.00393	.0158	.0642	.148	.455	1.074	1.642	2.706	3.841	5.412	6.635
2	.0201	.0404	.103	.211	.446	.713	1.386	2.408	3.219	4.605	5.991	7.824	9.210
3	.115	.185	.352	.584	1.005	1.424	2.366	3.665	4.642	6.251	7.815	9.837	11.341
4	.297	.429	.711	1.064	1.649	2.195	3.357	4.878	5.989	7.779	9.488	11.668	13.277
5	.554	.752	1.145	1.610	2.343	3.000	4.351	6.064	7.289	9.236	11.070	13.388	15.086
6	.872	1.134	1.635	2.204	3.070	3.828	5.348	7.231	8.558	10.645	12.592	15.033	16.812
7	1.239	1.564	2.167	2.833	3.822	4.671	6.346	8.383	9.803	12.017	14.067	16.622	18.475
8	1.646	2.032	2.733	3.490	4.594	5.527	7.344	9.524	11.030	13.362	15.507	18.168	20.090
9	2.088	2.532	3.325	4.168	5.380	6.393	8.343	10.656	12.242	14.684	16.919	19.679	21.666
10	2.558	3.059	3.940	4.865	6.179	7.267	9.342	11.781	13.442	15.987	18.307	21.161	23.209
11	3.053	3.609	4.575	5.578	6.989	8.148	10.341	12.899	14.631	17.275	19.675	22.618	24.725
12	3.571	4.178	5.226	6.304	7.807	9.034	11.340	14.011	15.812	18.549	21.026	24.054	26.217
13	4.107	4.765	5.892	7.042	8.634	9.926	12.340	15.119	16.985	19.812	22.362	25.472	27.688
14	4.660	5.368	6.571	7.790	9.467	10.821	13.339	16.222	18.151	21.064	23.685	26.873	29.141
15	5.229	5.985	7.261	8.547	10.307	11.721	14.339	17.322	19.311	22.307	24.996	28.259	30.578
16	5.812	6.614	7.962	9.312	11.152	12.624	15.338	18.418	20.465	23.542	26.296	29.633	32.000
17	6.408	7.255	8.672	10.085	12.002	13.531	16.338	19.511	21.615	24.769	27.587	30.995	33.409
18	7.015	7.906	9.390	10.865	12.857	14.440	17.338	20.601	22.760	25.989	28.869	32.346	34.805
19	7.633	8.567	10.117	11.651	13.716	15.352	18.338	21.689	23.900	27.204	30.144	33.687	36.191
20	8.260	9.237	10.851	12.443	14.578	16.266	19.337	22.775	25.038	28.412	31.410	35.020	37.566
21	8.897	9.915	11.591	13.240	15.445	17.182	20.337	23.858	26.171	29.615	32.671	36.343	38.932
22	9.542	10.600	12.338	14.041	16.314	18.101	21.337	24.939	27.301	30.813	33.924	37.659	40.289
23	10.196	11.293	13.091	14.848	17.187	19.021	22.337	26.018	28.429	32.007	35.172	38.968	41.638
24	10.856	11.992	13.848	15.659	18.062	19.943	23.337	27.096	29.553	33.196	36.415	40.270	42.980
25	11.524	12.697	14.611	16.473	18.940	20.867	24.337	28.172	30.675	34.382	37.652	41.566	44.314
26	12.198	13.409	15.379	17.292	19.820	21.792	25.336	29.246	31.795	35.563	38.885	42.856	45.642
27	12.879	14.125	16.151	18.114	20.703	22.719	26.336	30.319	32.912	36.741	40.113	44.140	46.963
28	13.565	14.847	16.928	18.939	21.588	23.647	27.336	31.391	34.027	37.916	41.337	45.419	48.278
29	14.256	15.574	17.708	19.768	22.475	24.577	28.336	32.461	35.139	39.087	42.557	46.693	49.588
30	14.953	16.306	18.493	20.599	23.364	25.508	29.336	33.530	36.250	40.256	43.773	47.962	50.892

* Table *E* is reprinted from Table III of Fisher's Statistical methods for research workers. Oliver & Boyd, Edinburgh, by kind permission of the author and publishers.

TABLE F.*—5 PER CENT (ROMAN TYPE) AND 1 PER CENT (**BOLD-FACED TYPE**) POINTS FOR THE DISTRIBUTION OF F

n_2	n_1 degrees of freedom (for greater variance)																								n_2
	1	2	3	4	5	6	7	8	9	10	11	12	14	16	20	24	30	40	50	75	100	200	500	∞	
1	161 **4,052**	200 **4,999**	216 **5,403**	225 **5,625**	230 **5,764**	234 **5,859**	237 **5,928**	239 **5,981**	241 **6,022**	242 **6,056**	243 **6,082**	244 **6,106**	245 **6,142**	246 **6,169**	248 **6,208**	249 **6,234**	250 **6,258**	251 **6,286**	252 **6,302**	253 **6,323**	253 **6,334**	254 **6,352**	254 **6,361**	254 **6,366**	1
2	18.51 **98.49**	19.00 **99.01**	19.16 **99.17**	19.25 **99.25**	19.30 **99.30**	19.33 **99.33**	19.36 **99.34**	19.37 **99.36**	19.38 **99.38**	19.39 **99.40**	19.40 **99.41**	19.41 **99.42**	19.42 **99.43**	19.43 **99.44**	19.44 **99.45**	19.45 **99.46**	19.46 **99.47**	19.47 **99.48**	19.47 **99.48**	19.48 **99.49**	19.49 **99.49**	19.49 **99.49**	19.50 **99.50**	19.50 **99.50**	2
3	10.13 **34.12**	9.55 **30.81**	9.28 **29.46**	9.12 **28.71**	9.01 **28.24**	8.94 **27.91**	8.88 **27.67**	8.84 **27.49**	8.81 **27.34**	8.78 **27.23**	8.76 **27.13**	8.74 **27.05**	8.71 **26.92**	8.69 **26.83**	8.66 **26.69**	8.64 **26.60**	8.62 **26.50**	8.60 **26.41**	8.58 **26.35**	8.57 **26.27**	8.56 **26.23**	8.54 **26.18**	8.54 **26.14**	8.53 **26.12**	3
4	7.71 **21.20**	6.94 **18.00**	6.59 **16.69**	6.39 **15.98**	6.26 **15.52**	6.16 **15.21**	6.09 **14.98**	6.04 **14.80**	6.00 **14.66**	5.96 **14.54**	5.93 **14.45**	5.91 **14.37**	5.87 **14.24**	5.84 **14.15**	5.80 **14.02**	5.77 **13.93**	5.74 **13.83**	5.71 **13.74**	5.70 **13.69**	5.68 **13.61**	5.66 **13.57**	5.65 **13.52**	5.64 **13.48**	5.63 **13.46**	4
5	6.61 **16.26**	5.79 **13.27**	5.41 **12.06**	5.19 **11.39**	5.05 **10.97**	4.95 **10.67**	4.88 **10.45**	4.82 **10.27**	4.78 **10.15**	4.74 **10.05**	4.70 **9.96**	4.68 **9.89**	4.64 **9.77**	4.60 **9.68**	4.56 **9.55**	4.53 **9.47**	4.50 **9.38**	4.46 **9.29**	4.44 **9.24**	4.42 **9.17**	4.40 **9.13**	4.38 **9.07**	4.37 **9.04**	3.46 **9.02**	5
6	5.99 **13.74**	5.14 **10.92**	4.76 **9.78**	4.53 **9.15**	4.39 **8.75**	4.28 **8.47**	4.21 **8.26**	4.15 **8.10**	4.10 **7.98**	4.06 **7.87**	4.03 **7.79**	4.00 **7.72**	3.96 **7.60**	3.92 **7.52**	3.87 **7.39**	3.84 **7.31**	3.81 **7.23**	3.77 **7.14**	3.75 **7.09**	3.72 **7.02**	3.71 **6.99**	3.69 **6.94**	3.68 **6.90**	3.67 **6.88**	6
7	5.59 **12.25**	4.74 **9.55**	4.35 **8.45**	4.12 **7.85**	3.97 **7.46**	3.87 **7.19**	3.79 **7.00**	3.73 **6.84**	3.68 **6.71**	3.63 **6.62**	3.60 **6.54**	3.57 **6.47**	3.52 **6.35**	3.49 **6.27**	3.44 **6.15**	3.41 **6.07**	3.38 **5.98**	3.34 **5.90**	3.32 **5.85**	3.29 **5.78**	3.28 **5.75**	3.25 **5.70**	3.24 **5.67**	3.23 **5.65**	7
8	5.32 **11.26**	4.46 **8.65**	4.07 **7.59**	3.84 **7.01**	3.69 **6.63**	3.58 **6.37**	3.50 **6.19**	3.44 **6.03**	3.39 **5.91**	3.34 **5.82**	3.31 **5.74**	3.28 **5.67**	3.23 **5.56**	3.20 **5.48**	3.15 **5.36**	3.12 **5.28**	3.08 **5.20**	3.05 **5.11**	3.03 **5.06**	3.00 **5.00**	2.98 **4.96**	2.96 **4.91**	2.94 **4.88**	2.93 **4.86**	8
9	5.12 **10.56**	4.26 **8.02**	3.86 **6.99**	3.63 **6.42**	3.48 **6.06**	3.37 **5.80**	3.29 **5.62**	3.23 **5.47**	3.18 **5.35**	3.13 **5.26**	3.10 **5.18**	3.07 **5.11**	3.02 **5.00**	2.98 **4.92**	2.93 **4.80**	2.90 **4.73**	2.86 **4.64**	2.82 **4.56**	2.80 **4.51**	2.77 **4.45**	2.76 **4.41**	2.73 **4.36**	2.72 **4.33**	2.71 **4.31**	9
10	4.96 **10.04**	4.10 **7.56**	3.71 **6.55**	3.48 **5.99**	3.33 **5.64**	3.22 **5.39**	3.14 **5.21**	3.07 **5.06**	3.02 **4.95**	2.97 **4.85**	2.94 **4.78**	2.91 **4.71**	2.86 **4.60**	2.82 **4.52**	2.77 **4.41**	2.74 **4.33**	2.70 **4.25**	2.67 **4.17**	2.64 **4.12**	2.61 **4.05**	2.59 **4.01**	2.56 **3.96**	2.55 **3.93**	2.54 **3.91**	10
11	4.84 **9.65**	3.98 **7.20**	3.59 **6.22**	3.36 **5.67**	3.20 **5.32**	3.09 **5.07**	3.01 **4.88**	2.95 **4.74**	2.90 **4.63**	2.86 **4.54**	2.82 **4.46**	2.79 **4.40**	2.74 **4.29**	2.70 **4.21**	2.65 **4.10**	2.61 **4.02**	2.57 **3.94**	2.53 **3.86**	2.50 **3.80**	2.47 **3.74**	2.45 **3.70**	2.42 **3.66**	2.41 **3.62**	2.40 **3.60**	11
12	4.75 **9.33**	3.88 **6.93**	3.49 **5.95**	3.26 **5.41**	3.11 **5.06**	3.00 **4.82**	2.92 **4.65**	2.85 **4.50**	2.80 **4.39**	2.76 **4.30**	2.72 **4.22**	2.69 **4.16**	2.64 **4.05**	2.60 **3.98**	2.54 **3.86**	2.50 **3.78**	2.46 **3.70**	2.42 **3.61**	2.40 **3.56**	2.36 **3.49**	2.35 **3.46**	2.32 **3.41**	2.31 **3.38**	2.30 **3.36**	12
14	4.60 **8.86**	3.74 **6.51**	3.34 **5.56**	3.11 **5.03**	2.96 **4.69**	2.85 **4.46**	2.77 **4.28**	2.70 **4.14**	2.65 **4.03**	2.60 **3.94**	2.56 **3.86**	2.53 **3.80**	2.48 **3.70**	2.44 **3.62**	2.39 **3.51**	2.35 **3.43**	2.31 **3.34**	2.27 **3.26**	2.24 **3.21**	2.21 **3.14**	2.19 **3.11**	2.16 **3.06**	2.14 **3.02**	2.13 **3.00**	14

TABLE F.*—5 PER CENT (ROMAN TYPE) AND 1 PER CENT (BOLD-FACED TYPE) POINTS FOR THE DISTRIBUTION OF F.—(*Continued*)

n_2	n_1 degrees of freedom (for greater variance)																							n_2
	1	2	3	4	5	6	7	8	9	10	11	12	14	16	20	24	30	40	50	75	100	200	500	∞	
17	4.45 **8.40**	3.59 **6.11**	3.20 **5.18**	2.96 **4.67**	2.81 **4.34**	2.70 **4.10**	2.62 **3.93**	2.55 **3.79**	2.50 **3.68**	2.45 **3.59**	2.41 **3.52**	2.38 **3.45**	2.33 **3.35**	2.29 **3.27**	2.23 **3.16**	2.19 **3.08**	2.15 **3.00**	2.11 **2.92**	2.08 **2.86**	2.04 **2.79**	2.02 **2.76**	1.99 **2.70**	1.97 **2.67**	1.96 **2.65**	17
20	4.35 **8.10**	3.49 **5.85**	3.10 **4.94**	2.87 **4.43**	2.71 **4.10**	2.60 **3.87**	2.52 **3.71**	2.45 **3.56**	2.40 **3.45**	2.35 **3.37**	2.31 **3.30**	2.28 **3.23**	2.23 **3.13**	2.18 **3.05**	2.12 **2.94**	2.08 **2.86**	2.04 **2.77**	1.99 **2.69**	1.96 **2.63**	1.92 **2.56**	1.90 **2.53**	1.87 **2.47**	1.85 **2.44**	1.84 **2.42**	20
24	4.26 **7.82**	3.40 **5.61**	3.01 **4.72**	2.78 **4.22**	2.62 **3.90**	2.51 **3.67**	2.43 **3.50**	2.36 **3.36**	2.30 **3.25**	2.26 **3.17**	2.22 **3.09**	2.18 **3.03**	2.13 **2.93**	2.09 **2.85**	2.02 **2.74**	1.98 **2.66**	1.94 **2.58**	1.89 **2.49**	1.86 **2.44**	1.82 **2.36**	1.80 **2.33**	1.76 **2.27**	1.74 **2.23**	1.73 **2.21**	24
30	4.17 **7.56**	3.32 **5.39**	2.92 **4.51**	2.69 **4.02**	2.53 **3.70**	2.42 **3.47**	2.34 **3.30**	2.27 **3.17**	2.21 **3.06**	2.16 **2.98**	2.12 **2.90**	2.09 **2.84**	2.04 **2.74**	1.99 **2.66**	1.93 **2.55**	1.89 **2.47**	1.84 **2.38**	1.79 **2.29**	1.76 **2.24**	1.72 **2.16**	1.69 **2.13**	1.66 **2.07**	1.64 **2.03**	1.62 **2.01**	30
40	4.08 **7.31**	3.23 **5.18**	2.84 **4.31**	2.61 **3.83**	2.45 **3.51**	2.34 **3.29**	2.25 **3.12**	2.18 **2.99**	2.12 **2.88**	2.07 **2.80**	2.04 **2.73**	2.00 **2.66**	1.95 **2.56**	1.90 **2.49**	1.84 **2.37**	1.79 **2.29**	1.74 **2.20**	1.69 **2.11**	1.66 **2.05**	1.61 **1.97**	1.59 **1.94**	1.55 **1.88**	1.53 **1.84**	1.51 **1.81**	40
50	4.03 **7.17**	3.18 **5.06**	2.79 **4.20**	2.56 **3.72**	2.40 **3.41**	2.29 **3.18**	2.20 **3.02**	2.13 **2.88**	2.07 **2.78**	2.02 **2.70**	1.98 **2.62**	1.95 **2.56**	1.90 **2.46**	1.85 **2.39**	1.78 **2.26**	1.74 **2.18**	1.69 **2.10**	1.63 **2.00**	1.60 **1.94**	1.55 **1.86**	1.52 **1.82**	1.48 **1.76**	1.46 **1.71**	1.44 **1.68**	50
70	3.98 **7.01**	3.13 **4.92**	2.74 **4.08**	2.50 **3.60**	2.35 **3.29**	2.23 **3.07**	2.14 **2.91**	2.07 **2.77**	2.01 **2.67**	1.97 **2.59**	1.93 **2.51**	1.89 **2.45**	1.84 **2.35**	1.79 **2.28**	1.72 **2.15**	1.67 **2.07**	1.62 **1.98**	1.56 **1.88**	1.53 **1.82**	1.47 **1.74**	1.45 **1.69**	1.40 **1.62**	1.37 **1.56**	1.35 **1.53**	70
100	3.94 **6.90**	3.09 **4.82**	2.70 **3.98**	2.46 **3.51**	2.30 **3.20**	2.19 **2.99**	2.10 **2.82**	2.03 **2.69**	1.97 **2.59**	1.92 **2.51**	1.88 **2.43**	1.85 **2.36**	1.79 **2.26**	1.75 **2.19**	1.68 **2.06**	1.63 **1.98**	1.57 **1.89**	1.51 **1.79**	1.48 **1.73**	1.42 **1.64**	1.39 **1.59**	1.34 **1.51**	1.30 **1.46**	1.28 **1.43**	100
150	3.91 **6.81**	3.06 **4.75**	2.67 **3.91**	2.43 **3.44**	2.27 **3.14**	2.16 **2.92**	2.07 **2.76**	2.00 **2.62**	1.94 **2.53**	1.89 **2.44**	1.85 **2.37**	1.82 **2.30**	1.76 **2.20**	1.71 **2.12**	1.64 **2.00**	1.59 **1.91**	1.54 **1.83**	1.47 **1.72**	1.44 **1.66**	1.37 **1.56**	1.34 **1.51**	1.29 **1.43**	1.25 **1.37**	1.22 **1.33**	150
200	3.89 **6.76**	3.04 **4.71**	2.65 **3.88**	2.41 **3.41**	2.26 **3.11**	2.14 **2.90**	2.05 **2.73**	1.98 **2.60**	1.92 **2.50**	1.87 **2.41**	1.83 **2.34**	1.80 **2.28**	1.74 **2.17**	1.69 **2.09**	1.62 **1.97**	1.57 **1.88**	1.52 **1.79**	1.45 **1.69**	1.42 **1.62**	1.35 **1.53**	1.32 **1.48**	1.26 **1.39**	1.22 **1.33**	1.19 **1.28**	200
400	3.86 **6.70**	3.02 **4.66**	2.62 **3.83**	2.39 **3.36**	2.23 **3.06**	2.12 **2.85**	2.03 **2.69**	1.96 **2.55**	1.90 **2.46**	1.85 **2.37**	1.81 **2.29**	1.78 **2.23**	1.72 **2.12**	1.67 **2.04**	1.60 **1.92**	1.54 **1.84**	1.49 **1.74**	1.42 **1.64**	1.38 **1.57**	1.32 **1.47**	1.28 **1.42**	1.22 **1.32**	1.16 **1.24**	1.13 **1.19**	400
1,000	3.85 **6.66**	3.00 **4.62**	2.61 **3.80**	2.38 **3.34**	2.22 **3.04**	2.10 **2.82**	2.02 **2.66**	1.95 **2.53**	1.89 **2.43**	1.84 **2.34**	1.80 **2.26**	1.76 **2.20**	1.70 **2.09**	1.65 **2.01**	1.58 **1.89**	1.53 **1.81**	1.47 **1.71**	1.41 **1.61**	1.36 **1.54**	1.30 **1.44**	1.26 **1.38**	1.19 **1.28**	1.13 **1.19**	1.08 **1.11**	1,000
∞	3.84 **6.64**	2.99 **4.60**	2.60 **3.78**	2.37 **3.32**	2.21 **3.02**	2.09 **2.80**	2.01 **2.64**	1.94 **2.51**	1.88 **2.41**	1.83 **2.32**	1.79 **2.24**	1.75 **2.18**	1.69 **2.07**	1.64 **1.99**	1.57 **1.87**	1.52 **1.79**	1.46 **1.69**	1.40 **1.59**	1.35 **1.52**	1.28 **1.41**	1.24 **1.36**	1.17 **1.25**	1.11 **1.15**	1.00 **1.00**	∞

* Reproduced from Snedecor, G. W. Statistical methods. Ames, Iowa: Collegiate, 1937. Pp. 174–177. By permission of the author.

TABLE G.—FUNCTIONS OF p, q, z, AND y, WHERE p AND q ARE PROPORTIONS ($p + q = 1.00$) AND z AND y ARE CONSTANTS OF THE UNIT NORMAL DISTRIBUTION CURVE*

	A	B	C	D	E	F	G	H	I	J	K	L	M	
p (or q)	pq	$\sqrt{pq}$	pq/y	$\sqrt{pq}/y$	p/y	y/p	zy/p	y	zy/q	y/q	q/y	$\sqrt{p/q}$	$\sqrt{q/p}$	q (or p)
.99	.0099	.0995−	.3715	3.733	37.15−	.02692	−.06262	.02665	6.2002	2.665	.3752	9.950	.1005	.01
.98	.0196	.1400	.4048	2.892	20.24	.04941	−.1015	.04842	4.9719	2.421	.4131	7.000	.1429	.02
.97	.0291	.1706	.4277	2.507	14.26	.07015	−.1319	.06804	4.2657	2.268	.4409	5.686	.1759	.03
.96	.0384	.1960	.4456	2.274	11.14	.08976	−.1571	.08617	3.7717	2.154	.4642	4.899	.2041	.04
.95	.0475	.2179	.4605	2.113	9.211	.1086	−.1786	.1031	3.3928	2.063	.4848	4.359	.2294	.05
.94	.0564	.2375−	.4735	1.994	7.891	.1267	−.1970	.1191	3.0868	1.985	.5037	3.958	.2526	.06
.93	.0651	.2551	.4848	1.900	6.926	.1444	−.2131	.1343	2.8307	1.918	.5213	3.645	.2743	.07
.92	.0736	.2713	.4951	1.825	6.188	.1616	−.2271	.1487	2.6110	1.858	.5381	3.391	.2949	.08
.91	.0819	.2862	.5043	1.762	5.604	.1785	−.2393	.1624	2.4191	1.804	.5542	3.180	.3145	.09
.90	.0900	.3000	.5128	1.709	5.128	.1950	−.2499	.1755	2.2491	1.755	.5698	3.000	.3333	.10
.89	.0979	.3129	.5206	1.664	4.733	.2113	−.2591	.1880	2.0966	1.709	.5850	2.844	.3516	.11
.88	.1056	.3250	.5279	1.625	4.399	.2273	−.2671	.2000	1.9587	1.667	.5999	2.708	.3693	.12
.87	.1131	.3363	.5346	1.590	4.112	.2432	−.2739	.2115	1.8330	1.627	.6145	2.587	.3865	.13
.86	.1204	.3470	.5409	1.559	3.864	.2588	−.2796	.2226	1.7175	1.590	.6290	2.478	.4035	.14
.85	.1275	.3571	.5468	1.532	3.646	.2743	−.2843	.2332	1.6110	1.554	.6433	2.380	.4201	.15
.84	.1344	.3666	.5524	1.507	3.452	.2896	−.2880	.2433	1.5123	1.521	.6576	2.291	.4365	.16
.83	.1411	.3756	.5576	1.484	3.280	.3049	−.2909	.2531	1.4203	1.489	.6718	2.210	.4525	.17
.82	.1476	.3842	.5625	1.464	3.125	.3200	−.2929	.2624	1.3344	1.458	.6860	2.134	.4685	.18
.81	.1539	.3923	.5671	1.446	2.985	.3350	−.2941	.2714	1.2538	1.428	.7002	2.065	.4844	.19
.80	.1600	.4000	.5715	1.429	2.858	.3500	−.2946	.2800	1.1781	1.400	.7144	2.000	.5000	.20
.79	.1659	.4073	.5756	1.413	2.741	.3648	−.2942	.2882	1.1067	1.372	.7287	1.940	.5156	.21
.78	.1716	.4142	.5796	1.399	2.634	.3796	−.2931	.2961	1.0393	1.346	.7430	1.883	.5311	.22
.77	.1771	.4208	.5832	1.386	2.536	.3943	−.2913	.3036	.9754	1.320	.7575	1.830	.5465	.23
.76	.1824	.4271	.5867	1.374	2.445	.4090	−.2889	.3109	.9149	1.295	.7720	1.780	.5620	.24
.75	.1875	.4330	.5900	1.363	2.360	.4237	−.2858	.3178	.8573	1.271	.7867	1.732	.5774	.25

* When p is less than .50, interchange p and q, as the headings of the first and last columns indicate.

TABLE G.—FUNCTIONS OF p, q, z, AND y, WHERE p AND q ARE PROPORTIONS ($p + q = 1.00$) AND z AND y ARE CONSTANTS OF THE UNIT NORMAL DISTRIBUTION CURVE.—(*Continued*)

	A	B	C	D	E	F	G	H	I	J	K	L	M	
p (or q)	pq	$\sqrt{pq}$	pq/y	$\sqrt{pq}/y$	p/y	y/p	zy/p	y	zy/q	y/q	q/y	$\sqrt{p/q}$	$\sqrt{q/p}$	q (or p)
.74	.1924	.4386	.5931	1.352	2.281	.4384	−.2820	.3244	.8026	1.248	.8016	1.687	.5928	.26
.73	.1971	.4440	.5961	1.343	2.208	.4529	−.2775	.3306	.7504	1.225	.8166	1.644	.6082	.27
.72	.2016	.4490	.5989	1.334	2.139	.4675	−.2725	.3366	.7006	1.202	.8318	1.604	.6236	.28
.71	.2059	.4538	.6015	1.326	2.074	.4822	−.2668	.3423	.6532	1.180	.8472	1.565	.6391	.29
.70	.2100	.4583	.6040	1.318	2.013	.4967	−.2605	.3477	.6078	1.159	.8628	1.528	.6547	.30
.69	.2139	.4625−	.6063	1.311	1.956	.5113	−.2535	.3528	.5643	1.138	.8787	1.492	.6703	.31
.68	.2176	.4665−	.6085	1.304	1.902	.5259	−.2460	.3576	.5227	1.118	.8949	1.458	.6860	.32
.67	.2211	.4702	.6106	1.298	1.850	.5405	−.2378	.3621	.4828	1.097	.9112	1.425	.7018	.33
.66	.2244	.4737	.6124	1.293	1.801	.5552	−.2290	.3664	.4445	1.078	.9279	1.393	.7178	.34
.65	.2275	.4770	.6142	1.288	1.755	.5698	−.2196	.3704	.4078	1.058	.9449	1.363	.7338	.35
.64	.2304	.4800	.6158	1.283	1.711	.5845	−.2095	.3741	.3725	1.039	.9623	1.333	.7500	.36
.63	.2331	.4828	.6174	1.279	1.669	.5993	−.1989	.3776	.3387	1.020	.9800	1.305	.7663	.37
.62	.2356	.4854	.6188	1.275	1.628	.6141	−.1876	.3808	.3061	1.002	.9980	1.277	.7829	.38
.61	.2379	.4877	.6200	1.271	1.590	.6290	−.1757	.3837	.2748	.9938	1.016	1.251	.7996	.39
.60	.2400	.4899	.6212	1.268	1.553	.6439	−.1631	.3863	.2447	.9659	1.035	1.225	.8165	.40
.59	.2419	.4918	.6223	1.265	1.518	.6589	−.1499	.3888	.2158	.9482	1.055	1.200	.8336	.41
.58	.2436	.4936	.6232	1.263	1.484	.6739	−.1361	.3909	.1879	.9307	1.074	1.175	.8510	.42
.57	.2451	.4951	.6240	1.260	1.451	.6891	−.1215	.3928	.1611	.9134	1.095	1.151	.8686	.43
.56	.2464	.4964	.6247	1.259	1.420	.7043	−.1063	.3944	.1353	.8964	1.116	1.128	.8864	.44
.55	.2475	.4975—	.6253	1.257	1.390	.7196	−.09043	.3958	.1105	.8796	1.137	1.106	.9045	.45
.54	.2484	.4984	.6258	1.256	1.360	.7351	−.07382	.3969	.0867	.8629	1.159	1.083	.9229	.46
.53	.2491	.4991	.6262	1.255	1.332	.7506	−.05650	.3978	.0637	.8464	1.181	1.062	.9417	.47
.52	.2496	.4996	.6264	1.254	1.305	.7662	−.03843	.3984	.0416	.8301	1.205	1.041	.9608	.48
.51	.2499	.4999	.6266	1.253	1.279	.7820	−.01960	.3988	.0204	.8139	1.229	1.020	.9802	.49
.50	.2500	.5000	.6267	1.253	1.253	.7979	−.00000	.3989	.0000	.7979	1.253	1.000	1.0000	.50

TABLE H.—CONVERSION OF A PEARSON *r* INTO A CORRESPONDING FISHER'S z COEFFICIENT*

r	z	*r*	z	*r*	z	*r*	z	*r*	z	*r*	z
.25†	.26	.40	.42	.55	.62	.70	.87	.85	1.26	.950	1.83
.26	.27	.41	.44	.56	.63	.71	.89	.86	1.29	.955	1.89
.27	.28	.42	.45	.57	.65	.72	.91	.87	1.33	.960	1.95
.28	.29	.43	.46	.58	.66	.73	.93	.88	1.38	.965	2.01
.29	.30	.44	.47	.59	.68	.74	.95	.89	1.42	.970	2.09
.30	.31	.45	.48	.60	.69	.75	.97	.90	1.47	.975	2.18
.31	.32	.46	.50	.61	.71	.76	1.00	.905	1.50	.980	2.30
.32	.33	.47	.51	.62	.73	.77	1.02	.910	1.53	.985	2.44
.33	.34	.48	.52	.63	.74	.78	1.05	.915	1.56	.990	2.65
.34	.35	.49	.54	.64	.76	.79	1.07	.920	1.59	.995	2.99
.35	.37	.50	.55	.65	.78	.80	1.10	.925	1.62		
.36	.38	.51	.56	.66	.79	.81	1.13	.930	1.66		
.37	.39	.52	.58	.67	.81	.82	1.16	.935	1.70		
.38	.40	.53	.59	.68	.83	.83	1.19	.940	1.74		
.39	.41	.54	.60	.69	.85	.84	1.22	.945	1.78		

* The values in this table were derived by interpolation from Table VB in Fisher's Statistical method for research workers, and are published by permission of the publisher, Oliver & Boyd, Edinburgh.

† For all values of *r* below .25, *r* = z.

TABLE J.—TRIGONOMETRIC FUNCTIONS[1]

ANGLE	SIN	COS	TAN	ANGLE	SIN	COS	TAN
0°	.000	1.000	.000	**45°**	.707	.707	1.000
1°	.018	.999	.018	**46°**	.719	.695	1.036
2°	.035	.999	.035	**47°**	.731	.682	1.072
3°	.052	.998	.052	**48°**	.743	.669	1.111
4°	.070	.997	.070	**49°**	.755	.656	1.150
5°	.087	.996	.087	**50°**	.766	.643	1.192
6°	.105	.994	.105	**51°**	.777	.629	1.235
7°	.122	.992	.123	**52°**	.788	.616	1.280
8°	.139	.990	.141	**53°**	.799	.602	1.327
9°	.156	.988	.158	**54°**	.809	.588	1.376
10°	.174	.985	.176	**55°**	.819	.574	1.428
11°	.191	.982	.194	**56°**	.829	.559	1.483
12°	.208	.978	.213	**57°**	.839	.545	1.540
13°	.225	.974	.231	**58°**	.848	.530	1.600
14°	.242	.970	.249	**59°**	.857	.515	1.664
15°	.259	.966	.268	**60°**	.866	.500	1.732
16°	.276	.961	.287	**61°**	.875	.485	1.804
17°	.292	.956	.306	**62°**	.883	.469	1.881
18°	.309	.951	.325	**63°**	.891	.454	1.963
19°	.326	.946	.344	**64°**	.899	.438	2.050
20°	.342	.940	.364	**65°**	.906	.423	2.144
21°	.358	.934	.384	**66°**	.914	.407	2.246
22°	.375	.927	.404	**67°**	.921	.391	2.356
23°	.391	.921	.424	**68°**	.927	.375	2.475
24°	.407	.914	.445	**69°**	.934	.358	2.605
25°	.423	.906	.466	**70°**	.940	.342	2.747
26°	.438	.899	.488	**71°**	.946	.326	2.904
27°	.454	.891	.510	**72°**	.951	.309	3.078
28°	.469	.883	.532	**73°**	.956	.292	3.271
29°	.485	.875	.554	**74°**	.961	.276	3.487
30°	.500	.866	.577	**75°**	.966	.259	3.732
31°	.515	.857	.601	**76°**	.970	.242	4.011
32°	.530	.848	.625	**77°**	.974	.225	4.331
33°	.545	.839	.649	**78°**	.978	.208	4.705
34°	.559	.829	.675	**79°**	.982	.191	5.145
35°	.574	.819	.700	**80°**	.985	.174	5.671
36°	.588	.809	.727	**81°**	.988	.156	6.314
37°	.602	.799	.754	**82°**	.990	.139	7.115
38°	.616	.788	.781	**83°**	.992	.122	8.144
39°	.629	.777	.810	**84°**	.994	.105	9.514
40°	.643	.766	.839	**85°**	.996	.087	11.430
41°	.656	.755	.869	**86°**	.997	.070	14.300
42°	.669	.743	.900	**87°**	.998	.052	19.081
43°	.682	731	.933	**88°**	.999	.035	28.636
44°	.695	.719	.966	**89°**	.999	.018	57.290

[1] From Smail, "College Algebra."

Table K.—Four-place Logarithms of Numbers[1]

N.	0	1	2	3	4	5	6	7	8	9
0	—	0000	3010	4771	6021	6990	7782	8451	9031	9542
1	0000	0414	0792	1139	1461	1761	2041	2304	2553	2788
2	3010	3222	3424	3617	3802	3979	4150	4314	4472	4624
3	4771	4914	5051	5185	5315	5441	5563	5682	5798	5911
4	6021	6128	6232	6335	6435	6532	6628	6721	6812	6902
5	6990	7076	7160	7243	7324	7404	7482	7559	7634	7709
6	7782	7853	7924	7993	8062	8129	8195	8261	8325	8388
7	8451	8513	8573	8633	8692	8751	8808	8865	8921	8976
8	9031	9085	9138	9191	9243	9294	9345	9395	9445	9494
9	9542	9590	9638	9685	9731	9777	9823	9868	9912	9956
10	0000	0043	0086	0128	0170	0212	0253	0294	0334	0374
11	0414	0453	0492	0531	0569	0607	0645	0682	0719	0755
12	0792	0828	0864	0899	0934	0969	1004	1038	1072	1106
13	1139	1173	1206	1239	1271	1303	1335	1367	1399	1430
14	1461	1492	1523	1553	1584	1614	1644	1673	1703	1732
15	1761	1790	1818	1847	1875	1903	1931	1959	1987	2014
16	2041	2068	2095	2122	2148	2175	2201	2227	2253	2279
17	2304	2330	2355	2380	2405	2430	2455	2480	2504	2529
18	2553	2577	2601	2625	2648	2672	2695	2718	2742	2765
19	2788	2810	2833	2856	2878	2900	2934	2945	2967	2989
20	3010	3032	3054	3075	3096	3118	3139	3160	3181	3201
21	3222	3243	3263	3284	3304	3324	3345	3365	3385	3404
22	3424	3444	3464	3483	3502	3522	3541	3560	3579	3598
23	3617	3636	3655	3674	3692	3711	3729	3747	3766	3784
24	3802	3820	3838	3856	3874	3892	3909	3927	3945	3962
25	3979	3997	4014	4031	4048	4065	4082	4099	4116	4133
26	4150	4166	4183	4200	4216	4232	4249	4265	4281	4298
27	4314	4330	4346	4362	4378	4393	4409	4425	4440	4456
28	4472	4487	4502	4518	4533	4548	4564	4579	4594	4609
29	4624	4639	4654	4669	4683	4698	4713	4728	4742	4757
30	4771	4786	4800	4814	4829	4843	4857	4871	4886	4900
31	4914	4928	4942	4955	4969	4983	4997	5011	5024	5038
32	5051	5065	5079	5092	5105	5119	5132	5145	5159	5172
33	5185	5198	5211	5224	5237	5250	5263	5276	5289	5302
34	5315	5328	5340	5353	5366	5378	5391	5403	5416	5428
35	5441	5453	5465	5478	5490	5502	5514	5527	5539	5551
36	5563	5575	5587	5599	5611	5623	5635	5647	5658	5670
37	5682	5694	5705	5717	5729	5740	5752	5763	5775	5786
38	5798	5809	5821	5832	5843	5855	5866	5877	5888	5899
39	5911	5922	5933	5944	5955	5966	5977	5988	5999	6010
40	6021	6031	6042	6053	6064	6075	6085	6096	6107	6117
41	6128	6138	6149	6160	6170	6180	6191	6201	6212	6222
42	6232	6243	6253	6263	6274	6284	6294	6304	6314	6325
43	6335	6345	6355	6365	6375	6385	6395	6405	6415	6425
44	6435	6444	6454	6464	6474	6484	6493	6503	6513	6522
45	6532	6542	6551	6561	6571	6580	6590	6599	6609	6618
46	6628	6637	6646	6656	6665	6675	6684	6693	6702	6712
47	6721	6730	6739	6749	6758	6767	6776	6785	6794	6803
48	6812	6821	6830	6839	6848	6857	6866	6875	6884	6893
49	6902	6911	6920	6928	6937	6946	6955	6964	6972	6981
50	6990	6998	7007	7016	7024	7033	7042	7050	7059	7067
N.	**0**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**

Prop. Parts

	22	**21**	**20**	**19**	**18**	**17**	**16**	**15**	**14**	**13**	**12**	**11**	**9**	**8**
1	2.2	2.1	2.0	1.9	1.8	1.7	1.6	1.5	1.4	1.3	1.2	1.1	0.9	0.8
2	4.4	4.2	4.0	3.8	3.6	3.4	3.2	3.0	2.8	2.6	2.4	2.2	1.8	1.6
3	6.6	6.3	6.0	5.7	5.4	5.1	4.8	4.5	4.2	3.9	3.6	3.3	2.7	2.4
4	8.8	8.4	8.0	7.6	7.2	6.8	6.4	6.0	5.6	5.2	4.8	4.4	3.6	3.2
5	11.0	10.5	10.0	9.5	9.0	8.5	8.0	7.5	7.0	6.5	6.0	5.5	4.5	4.0
6	13.2	12.6	12.0	11.4	10.8	10.2	9.6	9.0	8.4	7.8	7.2	6.6	5.4	4.8
7	15.4	14.7	14.0	13.3	12.6	11.9	11.2	10.5	9.8	9.1	8.4	7.7	6.3	5.6
8	17.6	16.8	16.0	15.2	14.4	13.6	12.8	12.0	11.2	10.4	9.6	8.8	7.2	6.4
9	19.8	18.9	18.0	17.1	16.2	15.3	14.4	13.5	12.6	11.7	10.8	9.9	8.1	7.2

[1] From Smail, "College Algebra."

TABLE K.—FOUR-PLACE LOGARITHMS OF NUMBERS.[1]—(*Continued*)

N.	0	1	2	3	4	5	6	7	8	9
50	6990	6998	7007	7016	7024	7033	7042	7050	7059	7067
51	7076	7084	7093	7101	7110	7118	7126	7135	7143	7152
52	7160	7168	7177	7185	7193	7202	7210	7218	7226	7235
53	7243	7251	7259	7267	7275	7284	7292	7300	7308	7316
54	7324	7332	7340	7348	7356	7364	7372	7380	7388	7396
55	7404	7412	7419	7427	7435	7443	7451	7459	7466	7474
56	7482	7490	7497	7505	7513	7520	7528	7536	7543	7551
57	7559	7566	7574	7582	7589	7597	7604	7612	7619	7627
58	7634	7642	7649	7657	7664	7672	7679	7686	7694	7701
59	7709	7716	7723	7731	7738	7745	7752	7760	7767	7774
60	7782	7789	7796	7803	7810	7818	7825	7832	7839	7846
61	7853	7860	7868	7875	7882	7889	7896	7903	7910	7917
62	7924	7931	7938	7945	7952	7959	7966	7973	7980	7987
63	7993	8000	8007	8014	8021	8028	8035	8041	8048	8055
64	8062	8069	8075	8082	8089	8096	8102	8109	8116	8122
65	8129	8136	8142	8149	8156	8162	8169	8176	8182	8189
66	8195	8202	8209	8215	8222	8228	8235	8241	8248	8254
67	8261	8267	8274	8280	8287	8293	8299	8306	8312	8319
68	8325	8331	8338	8344	8351	8357	8363	8370	8376	8382
69	8388	8395	8401	8407	8414	8420	8426	8432	8439	8445
70	8451	8457	8463	8470	8476	8482	8488	8494	8500	8506
71	8513	8519	8525	8531	8537	8543	8549	8555	8561	8567
72	8573	8579	8585	8591	8597	8603	8609	8615	8621	8627
73	8633	8639	8645	8651	8657	8663	8669	8675	8681	8686
74	8692	8698	8704	8710	8716	8722	8727	8733	8739	8745
75	8751	8756	8762	8768	8774	8779	8785	8791	8797	8802
76	8808	8814	8820	8825	8831	8837	8842	8848	8854	8859
77	8865	8871	8876	8882	8887	8893	8899	8904	8910	8915
78	8921	8927	8932	8938	8943	8949	8954	8960	8965	8971
79	8976	8982	8987	8993	8998	9004	9009	9015	9020	9025
80	9031	9036	9042	9047	9053	9058	9063	9069	9074	9079
81	9085	9090	9096	9101	9106	9112	9117	9122	9128	9133
82	9138	9143	9149	9154	9159	9165	9170	9175	9180	9186
83	9191	9196	9201	9206	9212	9217	9222	9227	9232	9238
84	9243	9248	9253	9258	9263	9269	9274	9279	9284	9289
85	9294	9299	9304	9309	9315	9320	9325	9330	9335	9340
86	9345	9350	9355	9360	9365	9370	9375	9380	9385	9390
87	9395	9400	9405	9410	9415	9420	9425	9430	9435	9440
88	9445	9450	9455	9460	9465	9469	9474	9479	9484	9489
89	9494	9499	9504	9509	9513	9518	9523	9528	9533	9538
90	9542	9547	9552	9557	9562	9566	9571	9576	9581	9586
91	9590	9595	9600	9605	9609	9614	9619	9624	9628	9633
92	9638	9643	9647	9652	9657	9661	9666	9671	9675	9680
93	9685	9689	9694	9699	9703	9708	9713	9717	9722	9727
94	9731	9736	9741	9745	9750	9754	9759	9763	9768	9773
95	9777	9782	9786	9791	9795	9800	9805	9809	9814	9818
96	9823	9827	9832	9836	9841	9845	9850	9854	9859	9863
97	9868	9872	9877	9881	9886	9890	9894	9899	9903	9908
98	9912	9917	9921	9926	9930	9934	9939	9943	9948	9952
99	9956	9961	9965	9969	9974	9978	9983	9987	9991	9996
100	0000	0004	0009	0013	0017	0022	0026	0030	0035	0039
N.	0	1	2	3	4	5	6	7	8	9

Prop. Parts

	9	8	7	6	5	4
1	0.9	0.8	0.7	0.6	0.5	0.4
2	1.8	1.6	1.4	1.2	1.0	0.8
3	2.7	2.4	2.1	1.8	1.5	1.2
4	3.6	3.2	2.8	2.4	2.0	1.6
5	4.5	4.0	3.5	3.0	2.5	2.0
6	5.4	4.8	4.2	3.6	3.0	2.4
7	6.3	5.6	4.9	4.2	3.5	2.8
8	7.2	6.4	5.6	4.8	4.0	3.2
9	8.1	7.2	6.3	5.4	4.5	3.6

[1] From Smail, "College Algebra."

APPENDIX C

A Glossary and Index of the More Important Symbols[1]

AD = average deviation from the mean (92).
a_{xy} = constant (Y intercept) in linear equation for regression of X on Y (399).
a_{yx} = constant in linear equation for regression of Y on X (399, 584).
a, b, c, d = observed frequencies in a 2 × 2 contingency table (231).
$a_x, b_x, \ldots, n_x$ = factor loadings of factors $A, B, \ldots, N$ in test X; also correlations between factors and test when factors are uncorrelated (517).
$a^2_x, b^2_x, \ldots, n^2_x$ = proportions of total variance in test X contributed by factors $A, B, \ldots, N$ (516).
b_{xy} = regression coefficient for linear dependence of X on Y (398).
b_{yx} = regression coefficient for linear dependence of Y on X (398, 584).
$b_{12.34\cdots m}$ = partial regression coefficient (428).
C = coefficient of contingency (343).
CV = coefficient of variation (118).
c = correction applied to a guessed mean (63).
c'_x = correction factor for X measured in class-interval units (102, 162).
cf = cumulative frequency (121).
cp = cumulative proportion (123).
D = difference between ranks (*e.g.*, $R_1 - R_2$) (311).
d = coefficient of determination (411).
df = degrees of freedom (185).
d_i = deviation of mean of a subsample I from the mean of a total sample (110, 580).
E = index of forecasting efficiency (410).
$E_{\infty x}$ = index of forecasting efficiency corrected for attenuation in the criterion (531).
e = base of natural logarithms (e = 2.718) (139).
e^2 = proportion of error variance in a set of measures (476).
F = Snedecor's F; ratio of one variance to another (232).
f = frequency; number of cases in a class or category (17).
f_e = expected frequency (52, 141, 275).
f_o = observed frequency (52).
f_x = frequency in a class interval on X (162).
GM = geometric mean (81).
HM = harmonic mean (84).
h^2_x = communality in test X; sum of proportions of common-factor variance (516).
i = size of class interval (63).
K^2 = coefficient of multiple nondetermination ($K^2 = 1 - R^2$) (432).
k = number of sets or subsamples in analysis of variance (238).
k = coefficient of alienation ($k = \sqrt{1 - r^2}$) (408).
k^2 = coefficient of nondetermination ($k^2 = 1 - r^2$) (411).
M = arithmetic mean (59).

[1] Numbers in parentheses refer to pages where more complete definitions can be found.

$\tilde{M}$ = mean of a population (a parameter) (176).
M' = a guessed mean (61).
M_c = mean of a column or Y array (394).
M_p = mean of one of two segments of a distribution whose proportion is p (324).
M_q = mean of one of two segments of a distribution whose proportion is q (324).
M_s = mean of an unweighted sum of measures (584).
M_{ws} = mean of a sum of weighted measures (453, 584).
M_∞ = mean of a set of true measures (475).
Mdn = median of a distribution (64).
Mo = mode of a distribution (69).
m = number of variables in a correlation problem, including dependent and one or more independent variables (434).
N = total number of cases (individuals or observations) in a sample (59).
N_P = number of cases in a total population (195).
n = (in analysis of variance) the number of cases in a subsample (238).
n = number of items or other equivalent parts in a test (493).
P = percentage (18).
P_{45} = 45th centile (124).
PE = probable error (104).
p = proportion; $p = f/N$ (19).
$\bar{p}$ (or $\bar{q}$) = mean of proportions (495).
$_cp$ = proportion of right responses corrected for chance success (550).
$\tilde{p}$ (or $\tilde{q}$) = a proportion true for a population (a parameter) (199).
$\tilde{p}_e$ (or $\tilde{q}_e$) = estimate of a population proportion (229).
p_l = proportion of a lower criterion group responding in a specified manner to a test item (503).
p_s = selection ratio (415).
p_u = proportion in an upper criterion group responding in a specified manner to an item (502).
Q = semi-interquartile range (90).
Q_1, Q_2, Q_3 = first, second, and third quartiles, respectively (90).
q = proportion complementary to p ($q = 1 - p$) (199).
R = number of items in a test answered correctly by an individual (533).
R^2 = coefficient of multiple determination (432).
R_{12} = coefficient of correlation in a sample of unrestricted range (349).
R_c = rank corrected for ties (313).
$R_{1.23\cdots m}$ = coefficient of multiple correlation (also denoted by R without subscripts) (426).
$_cR$ = a multiple R corrected for bias (434).
r = Pearson product-moment coefficient of correlation (157).
$\tilde{r}$ = correlation prevailing in a population (a parameter) (177).
r^2 = coefficient of determination (411).
r_b = biserial coefficient of correlation (324).
r_c = coefficient of correlation corrected for coarse grouping (360).
r_t = tetrachoric coefficient of correlation (334).
r_{cs} = correlation of a criterion C with an unweighted sum S of measures (462, 586).
r_{hh} = reliability estimate for halves of a test (492).
r_{ij} = correlation between items I and J (any two items); usually a phi coefficient (494).
$\bar{r}_{ij}$ = average correlation between items (494).

r_{it} = correlation between an item and total test score; usually a point-biserial r (494).
$\bar{r}_{it}$ = average correlation between item and total score (494).
r_{nn} = reliability of a homogeneous test of n units length (estimated by the Spearman-Brown formula) (493).
r_{pq} = correlation between a part and the remainder of a total (357).
r_{rr} = correlation between coefficients of correlation (224).
r_{tt} = reliability coefficient; proportion of true variance in a set of measures (476).
$r_{t\infty}$ = correlation between obtained and true measures; index of reliability ($r_{t\infty} = \sqrt{r_{tt}}$) (478).
r_{xy} = correlation between variables X and Y (also written r_{yx}) (157).
$r_{\infty x}$ = coefficient of correlation corrected for attenuation in one variable only (529).
$r_{\infty\omega}$ = coefficient of correlation corrected for attenuation in both variables (528).
r_{pbi} = point-biserial coefficient of correlation (329, 580).
$r_{c(ws)}$ = correlation between a criterion C and a weighted sum of variables (463, 587).
$r_{y(nx)}$ = correlation between Y and a test X increased homogeneously n times (527).
$r_{12.3}$ = a first-order partial coefficient of correlation (345).
$r_{12.34}$ = a second-order partial coefficient of correlation (346).
$r_{12}\sigma_1\sigma_2$ = covariance between X_1 and X_2 (454).
S_o = success ratio without the use of a certain selection device (415).
S_t = success ratio with the use of a selection test or other device (415).
SD = standard deviation (95).
SE = standard error (182).
t = ratio of deviation (from the mean) to σ (208).
V = variance in a sample distribution (98).
W = number of items in a test answered wrongly by an individual (533); also a weight applied to a test item (538).
w = a weight or multiplier (453).
X = obtained measurement or score (59).
X_c = a critical or minimum qualifying score in X (378, 384).
X_e = error contribution to an obtained measurement (474).
X_∞ = a true score in X (free from errors of measurement) (474).
X_{ba} = measurement in distribution B transformed into terms (same mean and standard deviation) of distribution A (572).
X_{ij} = a measurement in row I and column J (248).
x = a deviation from the mean ($x = X - M$).
x' = deviation of X from a guessed mean, in class-interval units (63).
x_s = deviation within a subsample from the subsample mean (239).
Y = an ordinate value (138).
Y' = a predicted Y value (399).
y = ordinate in a normal distribution curve whose total area equals 1.00 (140, 604).
y' = deviation of a predicted Y from the mean of Y ($y' = Y' - M_y$).
z = a standard measure, standard score, or deviate ($z = x/\sigma$).
$\mathbf{z}$ = Fisher's arc-tan function of r (212).
z_{ab} = a standard measure or deviate in units of difference between observed objects A and B (558).
$\alpha, \beta, \gamma, \delta$ = proportions within four cells of a 2×2 contingency table, corresponding to frequencies a, b, c, d (340).
$\beta_{12} = \beta_{12.34\cdots m}$.
$\beta_{12.34\cdots m}$ = beta coefficient; standard, partial regression coefficient (428).

η = correlation ratio (eta coefficient) (316).
ϕ = product-moment coefficient of correlation between two genuine dichotomies; Yule's phi (340, 582).
π = ratio of circumference of a circle to its diameter (138).
ρ = rank-difference coefficient of correlation; rho coefficient (312).
Σ = the sum of whatever follows it immediately (42).
Σ_1 = standard deviation in a distribution with unrestricted range (349).
Σx^2 = sum of squares (of deviations) (96).
Σxy = sum of products x times y; a sum of products of moments (157).
$\frac{\Sigma x_1 x_2}{N}$ = covariance between X_1 and X_2 (454, 585).
σ = standard deviation of a sample distribution (95).
${}_c\sigma$ = standard deviation with Sheppard's correction for errors of coarse grouping (108).
$\tilde{\sigma}$ = standard deviation of a population (a parameter) (176).
σ_b = standard error of a regression coefficient (407).
σ_c = standard deviation of a column or Y array (395).
$\sigma^2{}_e$ = amount of error variance in a set of measures (476).
σ_f = standard error of a frequency (202).
σ_i = standard deviation of a test item ($\sigma_i = \sqrt{p_i q_i}$) (489).
σ_M = an estimate of $\tilde{\sigma}_M$, derived from σ (184).
$\tilde{\sigma}_M$ = standard error of a mean computed from $\tilde{\sigma}$ (182).
$\sigma^2{}_m$ = total variance among means of subsamples (194).
σ_{Mdn} = standard error of a median (197).
σ_P = standard error of a percentage (202).
σ_p = standard error of a proportion (199).
σ_Q = standard error of a semi-interquartile range (198).
σ_r = standard error of a Pearson r (205).
σ_{r_o} = standard error or r, when $\tilde{r}$ is assumed to be zero (207).
σ_s = standard deviation of an unweighted sum of components (454, 585).
$\sigma^2{}_t$ = total amount of variance in a set of measures (489).
σ_w = standard error of weight W (538).
σ'_x = standard deviation of X measured in class-interval units (162).
$\sigma^2{}_{y'}$ = amount of predicted variance in Y (412).
σ_z = standard error of Fisher's z (212).
$\sigma^2{}_\infty$ = amount of true variance in a set of measures (476).
σ_η = standard error of a correlation ratio (318).
σ_ρ = standard error of a rho coefficient (313).
σ_σ = standard error of a standard deviation (198).
σ_{d_M} = standard error of a difference between means (213, 216).
σ_{d_p} = standard error of a difference between proportions (221).
σ_{d_r} = standard error of a difference between coefficients of correlation (223).
σ_{d_z} = standard error of a difference between two Fisher's z's (224).
σ_{d_σ} = standard error of a difference between standard deviations (222).
σ_{r_b} = standard error of a biserial r (325).
σ_{r_t} = standard error of a tetrachoric r (335).
σ_{ws} = standard deviation of a weighted sum of measures (457, 586).
$\sigma_{t\infty}$ = standard error of an obtained measurement (479).
σ_{xy} = standard error of estimate of X and Y (405).
σ_{yx} = standard error of estimate of Y from X (391, 405).

${}_c\sigma_{yx}$ = standard error of estimate corrected for bias (393, 407).
σ^2_{yx} = amount of nonpredicted variance in y (412).
$\tilde{\sigma}_{est}$ = estimate of $\tilde{\sigma}$ from sample statistics (184).
$\sigma_{1.23\cdots m}$ = standard error of multiple estimate (433).
${}_c\sigma_{1.23\cdots m}$ = standard error of multiple estimate corrected for bias (434).
$\sigma_{r_{12.34\ldots m}}$ = standard error of a partial r (347).
χ^2 = chi square (276).

Greek Letters Used

α = alpha.
β = beta.
γ = gamma.
δ = delta.
η = eta.
ϕ = phi.
π = pi.
ρ = rho.
σ = sigma.
Σ = capital sigma.
χ = chi.
ω = omega.

AUTHOR INDEX

SUBJECT INDEX

T